

WHAT IS GENOMICS AND WHY SHOULD I CARE?

David A Ray
Professor, Dept. of Biological Sciences
Texas Tech University

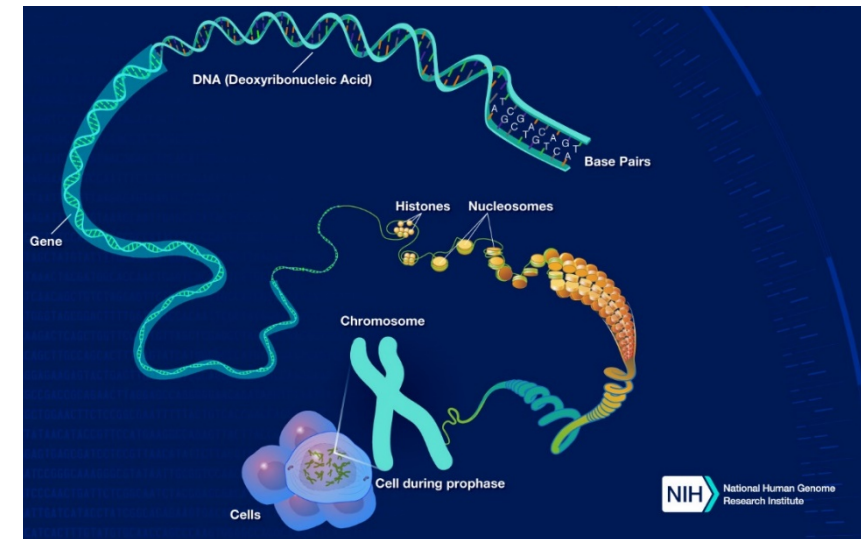


Genetics vs. Genomics

- The main difference:
 - Genetics typically examines the function and composition of single genes
 - Genomics typically addresses all functional aspects of a genome and their interrelationships

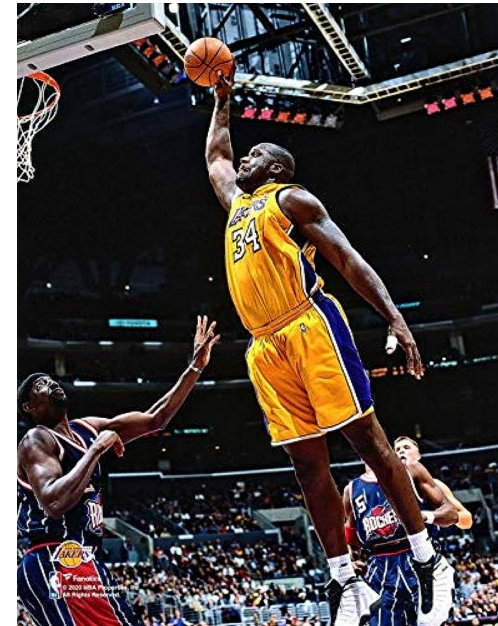
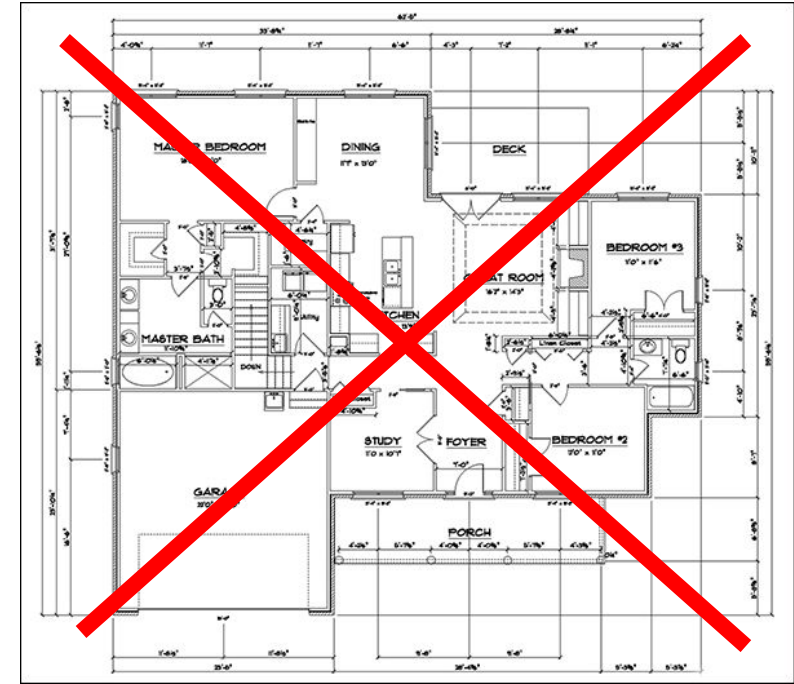
Genome

- Generic definition: Minimum nucleic acid complement that defines an organism/organelle/virus
 - Will vary depending on organism, organelle, or virus
 - Mostly DNA but some viruses have RNA-based genomes



Is the genome a blueprint?

- No. A blueprint is a scaled copy, a *genome is a chemical that interacts with the cellular and external environments to direct biochemical processes*
 - The DNA of our genome both facilitates and restricts the growth and function of a cell or multiple cells but does not **dictate** it.
 - Environmental impacts
 - Toxins, nutrition, exposure to disease
 - If someone were to clone you...?



Genomics

- Research in which robotics, automated sequencing, and advanced computational methods are utilized to *rapidly* and *efficiently* characterize genomes and their components
- Genomes vs. genome assemblies
- A **genome** is a physical thing in organisms that's *made up of DNA*
- A **genome assembly** is an approximation of a genome *made up of bits and bytes*
- While genomes can be investigated directly, we often use assemblies as their proxies.



Genomics

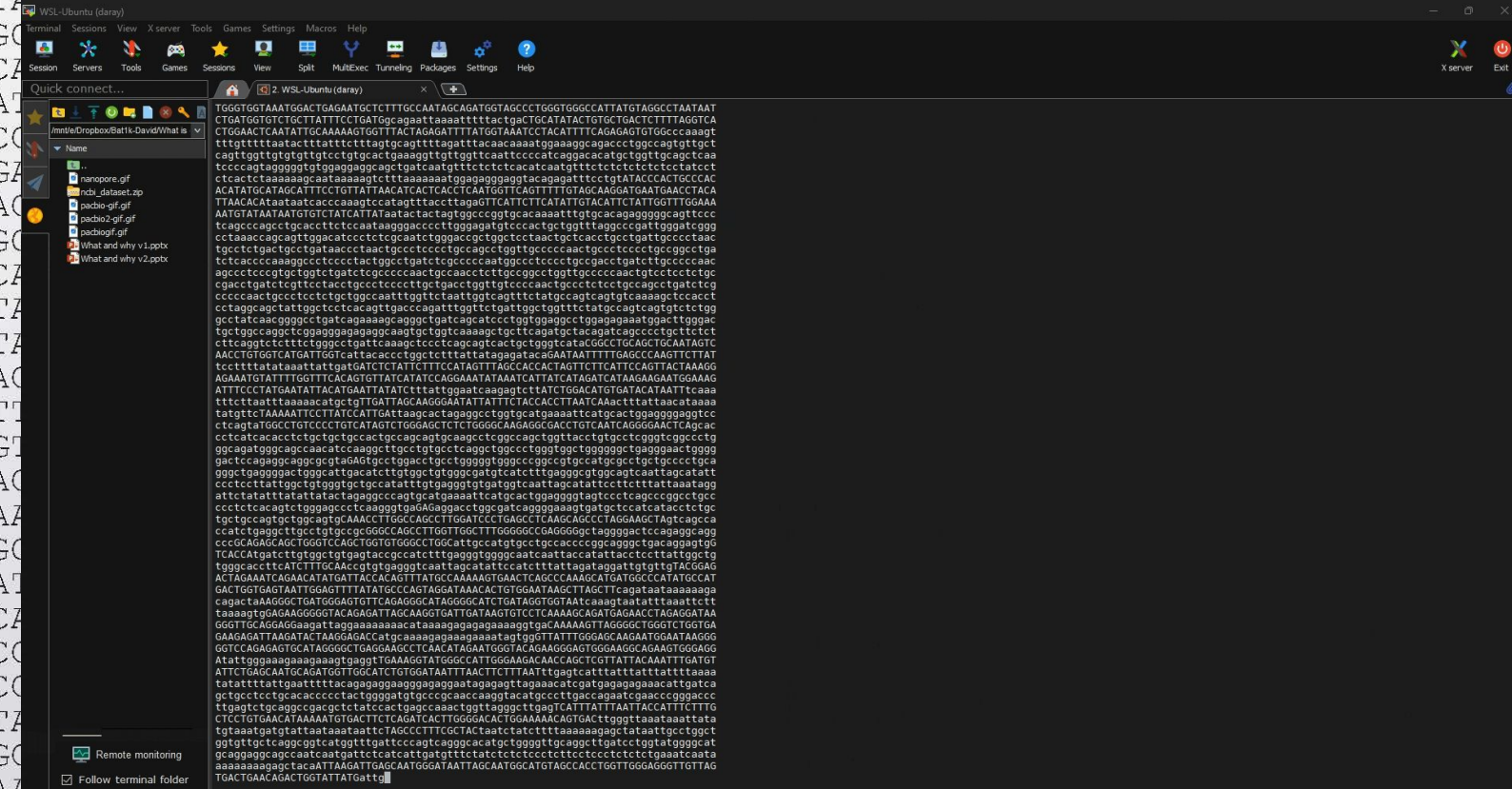
- Many different subfields
 - Applied genomics
 - Population genomics
 - Regulatory genomics
 - Structural genomics
 - Bioinformatics
 - Metagenomics
 - Functional genomics
 - Disease genomics
 - Comparative genomics
 - Phylogenomics



TATTGAAATTTTCAAAAATTTCTTACTTTTTTTTTTTGGATGGACGCAAGAAAGTTTAAATAATCATATACATGGCCATTACCACCATATAT
ATCCATATCTAATCTTACTTATATGTTGTGGAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAACCTTCTCTTTGGAACCTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCCTCGCGCCGCACACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATC
ATGCGATTA

What do genomes look like and how do we make sense of them?

CTGCATAAC TAATATACC TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGC
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAAATAAATGCTGATCCCAAATTTGCTCAAAGGA
CGATTTGCCGTTGGACGGTTCATTATGTCACAATPGATCCTTCTGTGTGCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTT
ACTCTTTTTCTAAAGAACTTGCACCGGAAAGGTTTTGCCAGTGCCTCCTCGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACC
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATA
CAAGCAAAAATTAATGCGTATTACGGTCTGTTGCAGAACATATGTTGGGTAAACAATGGCGGATGGATCAGGCTGCCTCTGTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTA AAAAACCATGA
AGCTTTGTTATTGCGAACACCCTTGTGTA

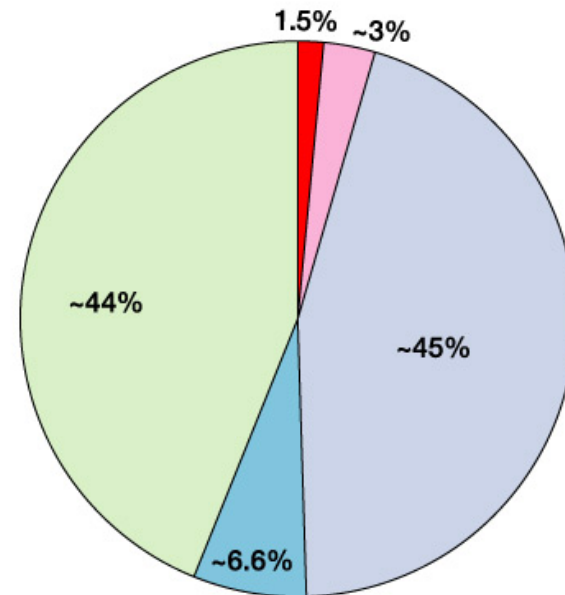


AGCTGCAAATGTTTTAGCTGCCACGTACG
TCATGAACGTTTATTATGCCAGATATCAC
CTAGTACTAGTTGAAGAGTCTCTCGCCAA
ATTCACAAGAGACTACTTAACAACATCTC
TAAGAGTCTTGAAGGCTGTGAAATTAATG
GAGTCTCAAGCTTCTTGGGATAAACTTAC
TGGTTCCCGTTTTGACCGGAGCTGGCTGGG
AAGCCCTTGCCAATGAGTTCTACAAGGTC
TTGGGCAGCTGTCTATATGAATTATAAGT
GCATCACAAAATACGCAATAATAACGAGT
TTTTCAATGTAAGAGATTTTCGATTATCCA
CCTATTCTTGACATGATATGACTACCATT
TTGGCAAGTTGCCAACTGACGAGATGCAG
AACCCTTTGTCTACTGATTAATTTTGTAC
AAATTCCGATGGACAAGAAGATAGGAAAA
ATTTTCAAGTTAGACAAGGACAAAATCAG
TCCATTAATCTCTGTTCTCTCTTACTTA
CCATAGAGAAGATCTTTCGGTTTCGAAGAC
GCGGCTCTTCAAAAAGATTGAACTCTCGC
TATAAGTCATCTCAGAGTAATATAACTAC
CTCATTCTGGAAGAAAATCTATTATGAAT
AGGACTTGAAGCCCGTCGAAAAAGAAAGG
ACTTGGCAAATTCAGCTACAGGTCTACA

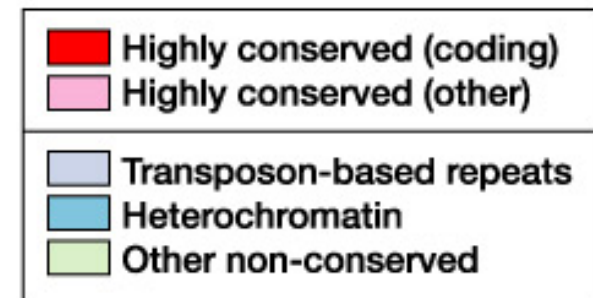
Genome Components

- Be aware
 - that many parts of a genome may belong to several categories
 - that I am primarily talking about complex, eukaryotic genomes
- Non-coding DNA
- Repetitive DNA
- Coding DNA

The human genome as an example

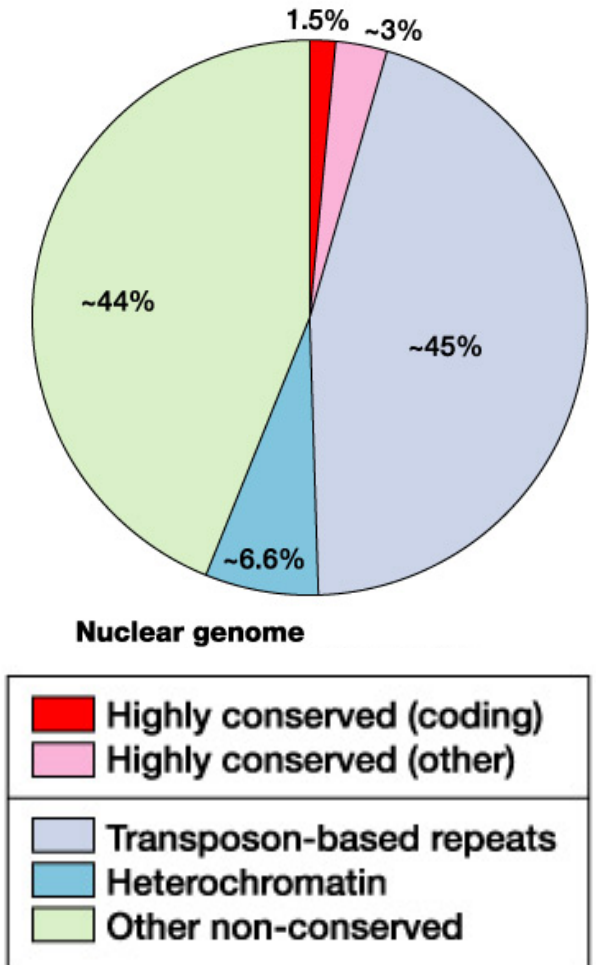


Nuclear genome

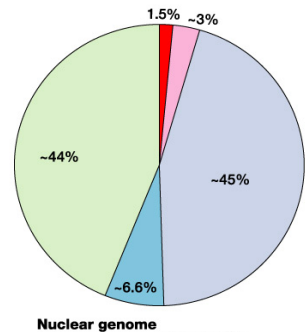


Genes vs. alleles vs. loci

- **Gene:** “Region of DNA that controls a discrete hereditary characteristic, often (but not always) corresponding to a single protein or RNA. This definition includes the entire functional unit, encompassing coding DNA sequences, non-coding regulatory DNA sequences, and introns.”
- **Allele:** “One of a set of alternative forms of a gene.”
- **Locus:** “The position of a gene on a chromosome. Different alleles of the same gene all occupy the same locus.”




TATTGAATTTTCAAAAATTTCTTACTTTTTTTTTTTGGATGGACGCAAGAAAGTTTAAATCATATTACATTTACCATTACCACCATATA
 ATCCATATCTAATCTTACT**TTATA**TGTTGTGGAAATGTAAAGAGCCCCATTAATCTTAGCCTAAAAAACCTTCTCTTTGGAACTTTC
 AATACGCTTAACTGCTCATTGCTATATTGAAGTA**CGG**ATTAGAAGCCG**CCG**AG**CGG**GCGACAGCCCT**CCGA****CGG**AAGACTCTCCTC
 GCGTCCCTCGTCTTCACCGGTCGCGTTCCCTGAAACGCAGATGTGCCT**CGC**GCCGCACTGCT**CCG**AACAATAAAGATTCTACAATACT
 TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGG**CCCCA**CAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGA
 ATGCGATTAGTTTTTTAGCCTTATTTCT**TGGGG**TAATTAATCAGCGAAGCGATGATTTTTGATCTATTAAACAGATA**TATAA**ATGG
 CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAA
 TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATA**ATGACTAAATCTCATT****CAGAAGAAGTGATTGTACCTGAGTTCT**
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACC
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTG
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGG
CGATTTGCCGTTGGACGGTTCATTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGT
ACTCTTTTTCTAAAGAACTTGCACCGGAAAGTTTTGCCAGTGTCTCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTAC
GGCAGTGGATTGTCTTCTTCGGCCGCATTTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTAAACAATGGCGGTATGGATCAGGCTGCCTCTGT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTA AAAAACCATG
AGCTTTGTTATTGCGAACACCCTTGTGATCTAACAGTTTTGAAACCGCCCAACCAACTATAATTTAAGAGTGGTAGAAGTC
AGCTGCAAAATGTTTTAGCTGCCACGTACGGTGTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTTATTATGCCAGATATCACAACTTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAC
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTGCGACAATCCTTGAATTGTTCTCGCG
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATC
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTCCCGTTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACC
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTTACTTTGTTTCCAGAACAACTTCTCATTTTTTTTCTACTCATAACT
 GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
 TTTTCAATGTAAGAGATTTTCGATTATCCACAACTTTTAAAACACAGGGACAAAATTTCTTGATATGCTTTCAACCGCTGCGTTTTGC
 CCTATTCTTGACATGATATGACTACCATTTTGTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAC
 TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTTGTCTTTTTTTTTTTCCGGGGACTCTAC
 AA**CCTTTTGT**CCTACTGATTA**TTTTGTACT**GAAATTT**GGACAAT**TCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGAC
 AAATTCAGATGGACAAGAAGATAGGAAAAAAAAGCTTTTACCAGATTTCCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
 ATTTTCAAGTTAGA**CAAGGAC**AAAATCAGGACAAATTTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCC
 TCCATTAAATCTCTGTTCTCTTACTTATATGATGATTAGGTATCATCTG**TATAA**AACTCCTTTCTTAAATTTCACTCTAAAGCAT
 CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAAATAGGAGGGAATA**ATGCCAGACAATCTATCATTACAT**
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAAC TTATGGAACTTTCCAATGAGACCTTTGCGCCAATAATGTGGATTGGA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAGTAAGCTCAGAAAACCTCAAT
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCAATTTATGTAC
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTGGTCTGGTACAATTAATGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTCAGCTACAGGTCTACAAC TGGGTC TAAATTTGGTGGCAGTGTGGATAACAATTTGGATTGGGTACGGTTTC



- Highly conserved (coding)
- Highly conserved (other)
- Transposon-based repeats
- Heterochromatin
- Other non-conserved


TATTGAATTTTCAAAAATTTCTTACTTTTGTGATGGACGCAAGGTTTAAATCATATGCATGGCATTACCACCATATA
 ATCCATATCTAATCTTAC**TTATA**TGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAACCTTCTCTTTGGAACTTTC
 AATACGCTTAACTGCTCATTGCTATATTGAAGTA**CGG**ATTAGAAGCCG**CCG**AG**CGG**GCGACAGCCCT**CCGA****CGG**AAGACTCTCCTC
 GCGTCCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCT**CGC**GCCGCACTGCT**CGG**AACAATAAAGATTCTACAATACT
 TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGG**CCCCA**CAAACCTTCAAATTAACGATCAAATTAACAACCATAGGA
 ATGCGATTAGTTTTTTAGCCTTATTT**TGGGG**TAATTAATCAGCGAATGATGATTTTTGATCTTAAACAGATA**TATAA**ATGG
 CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTTTATTCAAATGTCTAAAAGTATCAACAAAAA
 TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATA**ATGACTAAAT**CAATTCAGAAGAA**CGATTGTACCTGAGTT**C
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGATTTATAAGCGTTATGATGCTAAACC
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACCTCGGTTTACCTTTAGCTATTG
GATATCGCTTTCGGCTCAAACTTTTGAACGACATCCATCCATTACCTTAATAAATGCATCCCAATTTGCTCAAAGG
 CGTATTTGATCCTTCTGTGTCGGACTGGTCTA
 AC**GAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTG**
 GG**ATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAA**
 CA**TCGTTGCAGAACATTATGTTGGTGTAAACAATG**
 GT**GAGTTCAAACCGCAGTTGAAGGCTACTCCGTTT**
 AG**TGTATCTAACAGTTTGAAACCGCCCCAACCAA**
 AG**ACGGTGTGTTTTACTTTCTGGAAAAGAAGGAT**
 TC**CACAACATTTCCACACCCTGGAACGGCGATATT**
 CTAGTACTAGTTGAAGAGTCTCT**GCCAATAAGAAACAGGGCTTTAGTGTTGACGATGT**
 ATTCACAAGAGACTACTTAACAAC**CTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGG**TAAGCATGTGTATTCTGAAT
 TAAGAGTCTTGAAGGCTGTGAAATTA**TGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTT**TAAGCAATTTGGTGCCTTGATC
 GAGTCTCAAGCTTCTTGCGATAAACTT**ACGAATGTTCTTGTCAGAGATTGACAAAATTTGT**CCATTGCTTTGTCAAATGGATC
 TGGTTC**CCGTTTGACC**GGAGCTGGCTGG**STGGTGTACTGTTCACTTGGTTCAGGGGGCC**AAATGGCAACATAGAAAAGGTAF
 AAGCCCTTGCCAATGAGTCTACAAGGT**CAAGTACCCTAAGATCACTGATGCTGAGCTAG**AAATGCTATCATCGTCTCTAAACCF
TTGGGCAGCTGTCTATATGAATTATAAGTATCTTCTTTTTTTACTTTGTTT**CAGAACA**ATTTCTCATTTTTTTTCTACTCATAACT
 GCATCACAAAATACGCAATAA**TAAACGAGTAGT**CACTTTTTATAGTT**CATACATGCTT**CACTACTTAATAAATGATTGTATGATA
 TTTTCAATGTAAGAGATTT**CGATTATCCACAAAC**TTAA**AAACACAGGGACAAAAT**CTGATATGCTTTCAACCGCTGCGTTTTG
 CCTATTCTTGACATGATATGACTACCATTTT**GTTA**TGTACGTGGGGCAGTTGACGTTATCATATGTCAAAGTCATTTGCGAAC
 TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAG**CAATTGCCGTCTTGAACT**TTGTCCTTTTTTTTTTCCGGGGACTCTAC
 AA**CCTTTTGT**CTTACTGATTAA**TTTTGTACT**TGAATTT**CAACAAT**TCAGATTTTACAGACAAGCGCGAGGAGGAAAAGAAATGAC
 AAATTC**CGATGGACAAGAAGATAGGAAAAA**AAAAAGCTT**CACCGATTTC**TAACCGGAAAAAAGTCGTATGACATCAGAATGA
 ATTTTCAAGTTAGA**CAAGGAC**AAAATCAGGACAAATTGTAGATATAA**TAAAT**ATTTGATT**CAGCGCCAATTTGCC**CTTTTCCA
 TCCATTAAATCTCTGTTCTCTTACTTATATGATGATTAG**ATCATCTGTATAA**ACTCCTTTCTTAATTTCACTCTAAAGCA
 CCATAGAGAAGATCTTTCGGTTCGAAGACATT**CCTACGCATAA**AGAATAGGAGGGAA**ATAATGCCAGACAATCTATCATTACAT**
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCT**CAATGAGACCTTTGCGCCA**ATAATGTGGATT**TGGAAAA**
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAGTAAGCTCAGAAAACCTCAAT
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTAC
 AGGACTTGAAGCCCGTCGAAAAGAAAGGCGGGT**TGGT**CCTGGTACAAT**TATTGTTACTTCTGGCTTGCTGAATGTTT**CAATATC
 ACTTGGCAAAT**TGCAGCTACAGGTCTACA**ACTGGGTCTAAAT**TGGTGGCAGTGTGGATAACAATTTGGATTGGGTACGGTTTC**

Genes

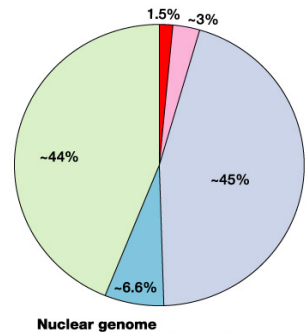


Encode proteins

Regulatory motifs



Control gene expression



- Highly conserved (coding)
- Highly conserved (other)
- Transposon-based repeats
- Heterochromatin
- Other non-conserved

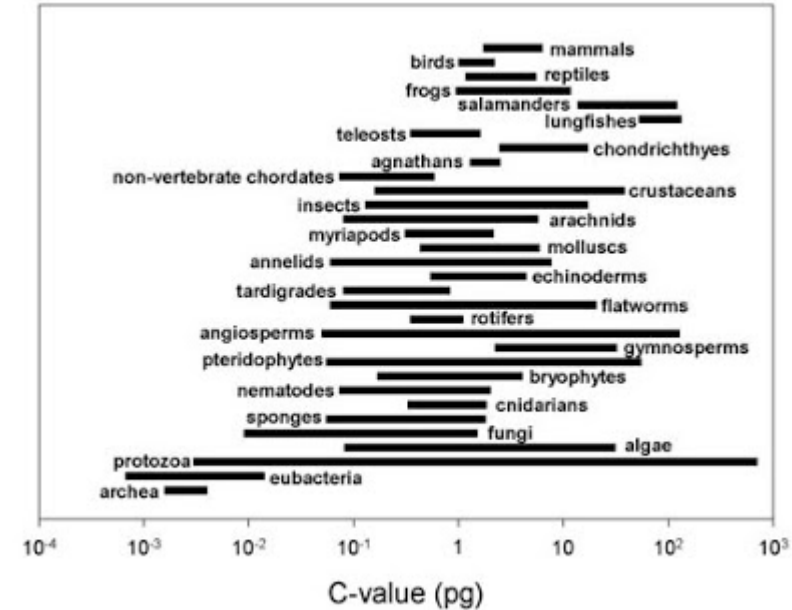
Sizes and Organization of Genomes

- There is no “one” genomic sequence for a species
- The differences in DNA sequence among members of the same species are generally relatively few, but they can be very important.
- **Variant:** any individual variation in the genomic content when compared to the population as a whole. Could be a SNP, indel, CNV, etc.

Genome Sizes

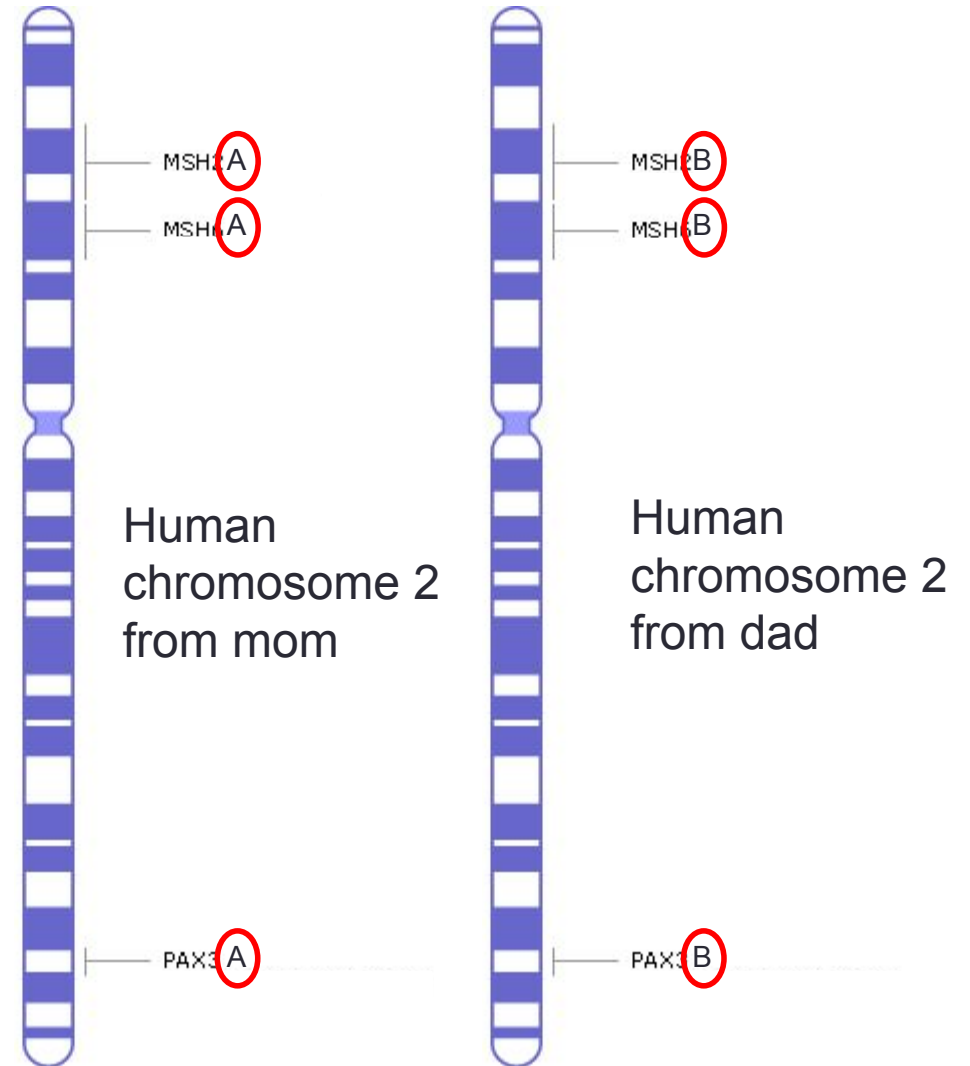
- **Genome Size and Organism Complexity**

- Very rough, imperfect correlation
- C-value – the characteristic value of haploid DNA content per nucleus; aka how much DNA is there in a genome?
- *C-value paradox*: the observation that genome size does not correlate with organismal complexity
- Taft et al. (2007) - complexity can be "broadly defined as the number and different types of cells, and the degree of cellular organization."
- *The onion test is a simple reality check for anyone who thinks they have come up with a universal function for non-coding DNA. Whatever your proposed function, ask yourself this question: Can I explain why an onion needs about five times more non-coding DNA for this function than a human?*



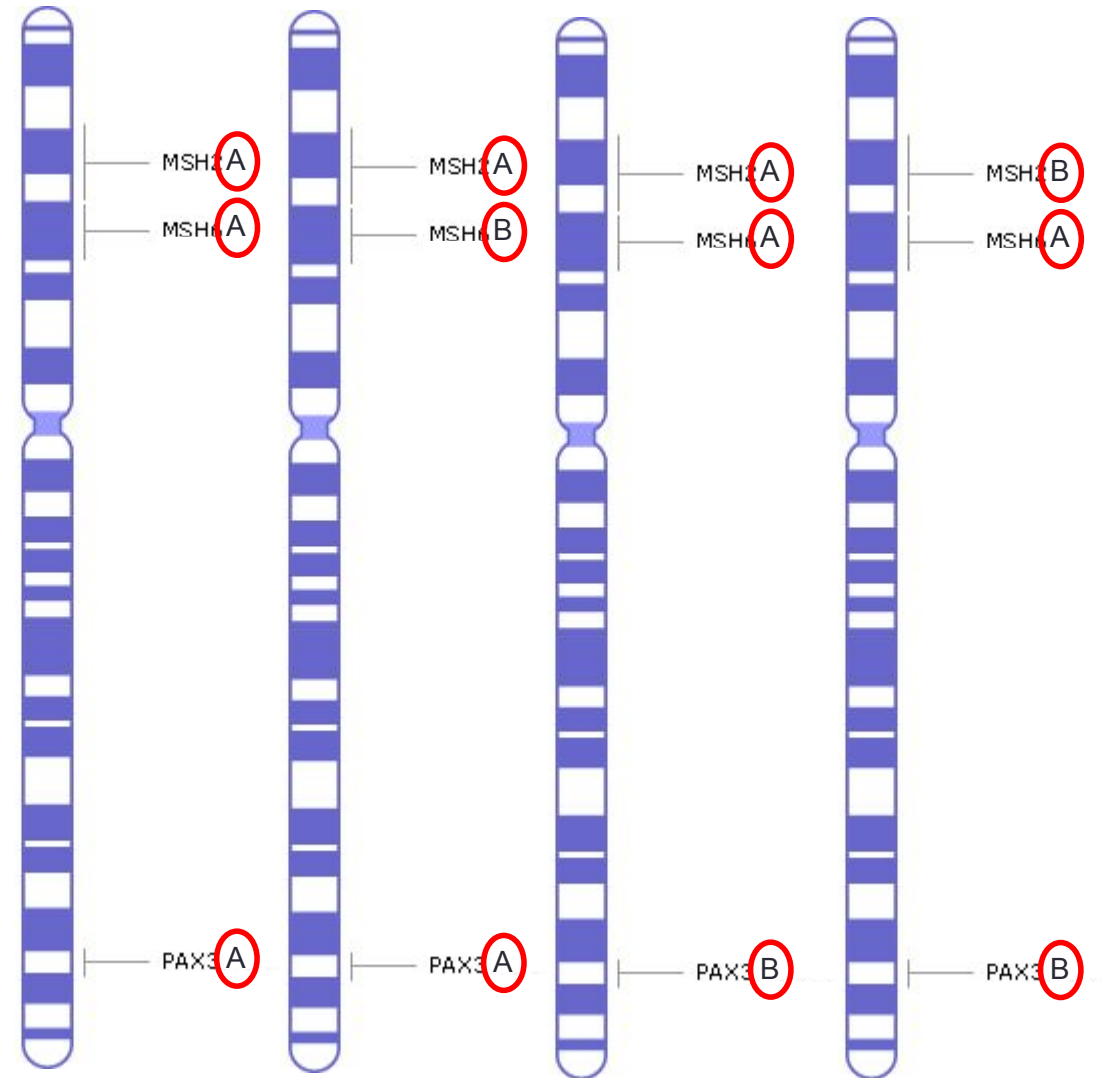
Chromosomes and genome maps

- DNA in most eukarotes is organized into linear chromosomes
 - Having *genes* all lined up on a chromosome suggests that the *alleles* would assort together, not independently.... but as a **linkage group**
 - Most *alleles* on a chromosome did assort independently, not as linkage groups.... how?
 - Some mechanism to allow neighboring *alleles* to assort independently must exist.



Chromosomes and genome maps

- DNA in most eukarotes is organized into linear chromosomes
 - Having *genes* all lined up on a chromosome suggests that the *alleles* would assort together, not independently.... but as a **linkage group**
 - Most *alleles* on a chromosome did assort independently, not as linkage groups.... how?
 - Some mechanism to allow neighboring *alleles* to assort independently must exist.



Possible chromosomes in gametes

Genetic Mapping

- **Crossing-over and recombination**
- Physical distance between loci influences recombination rates between loci
- Recombination fraction is a measure of distance

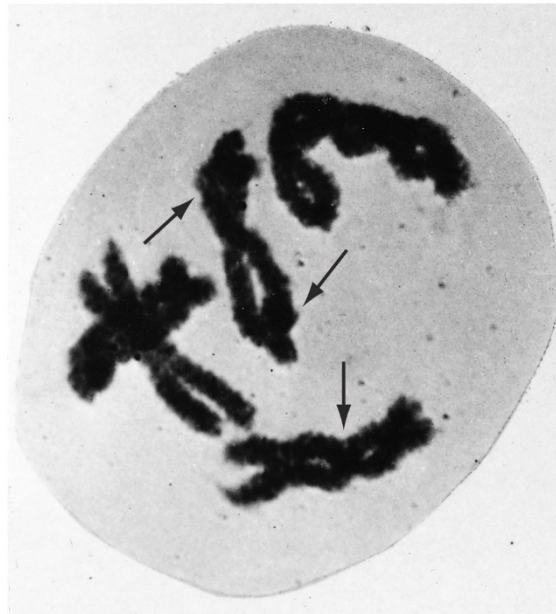
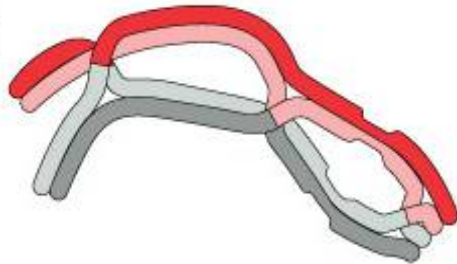
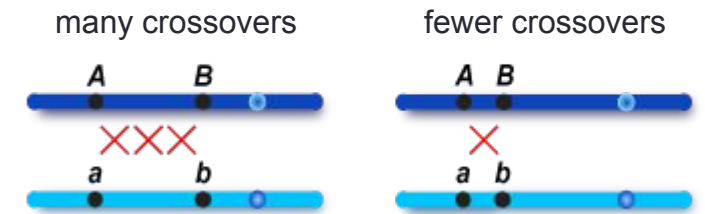


Figure 10-6 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

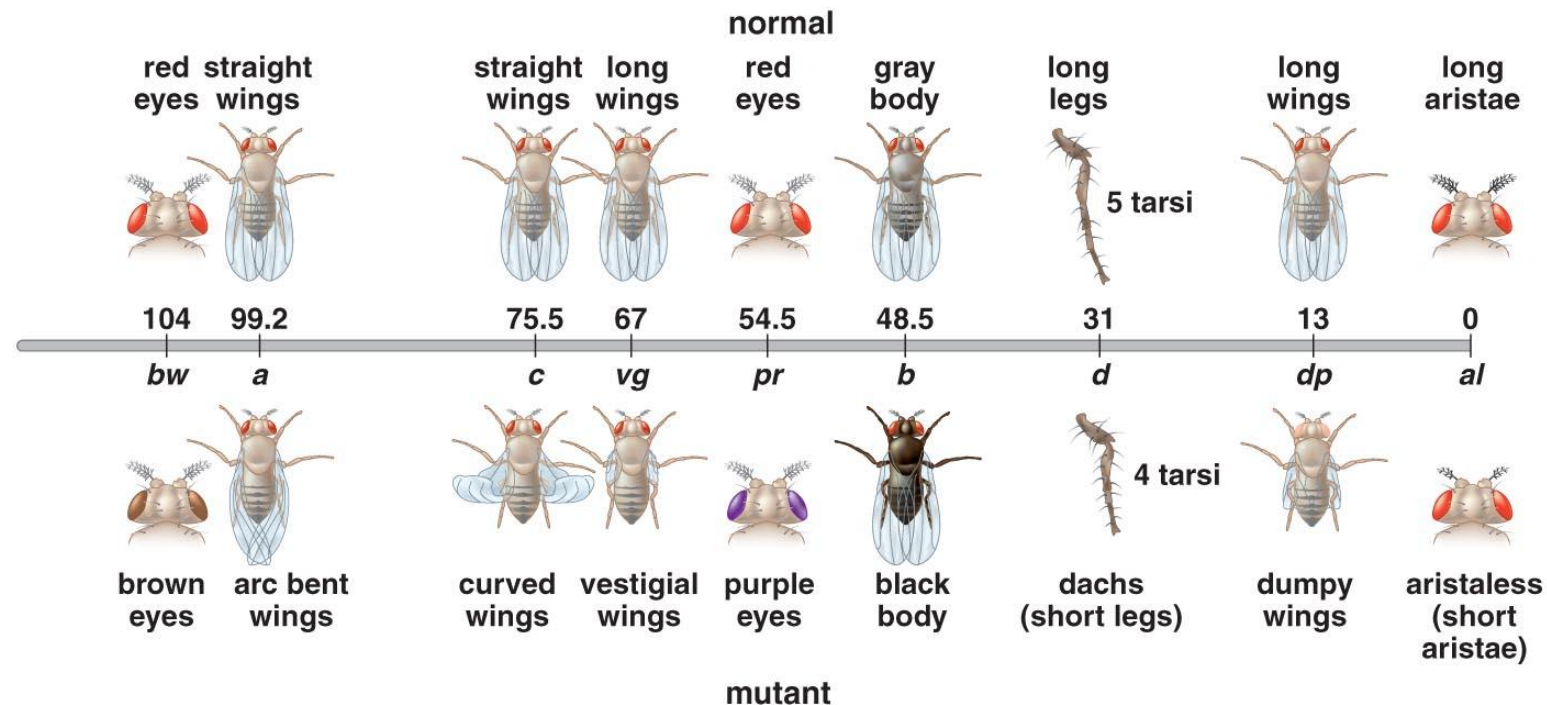
Distance and frequency of recombination between two points



© Tomáš Urban 2013

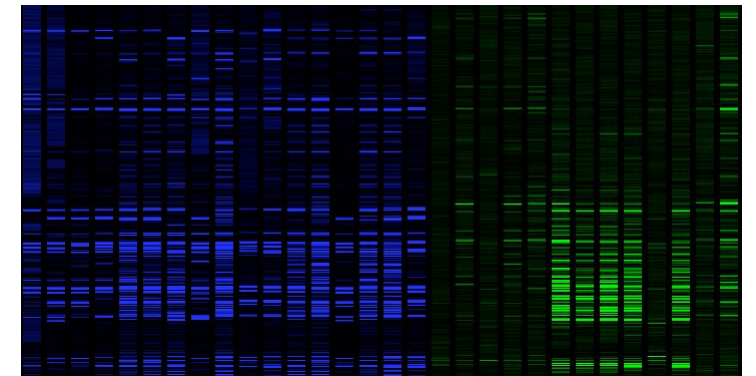
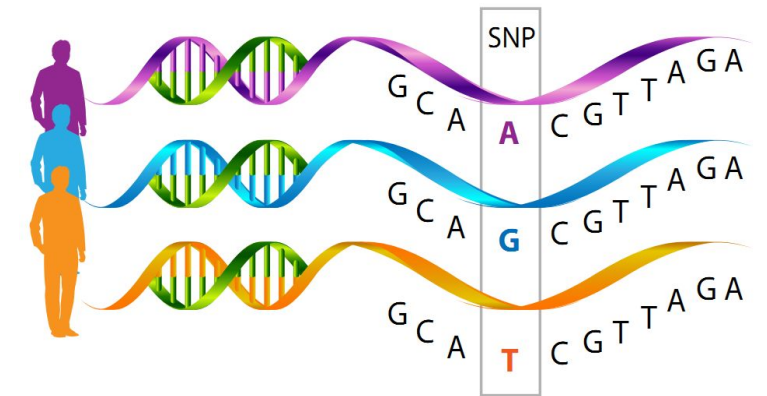
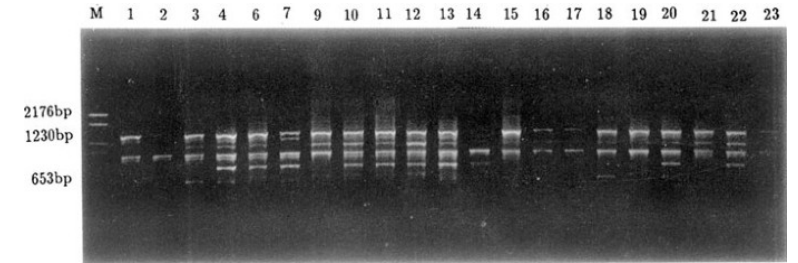
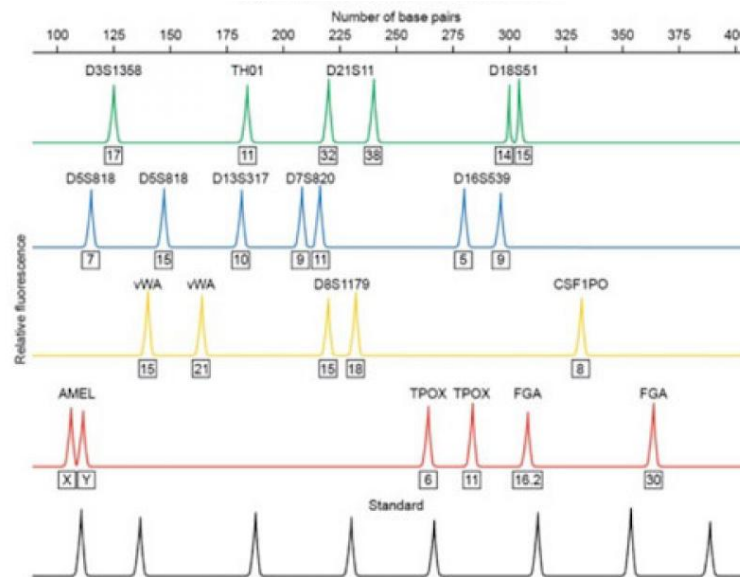
Genetic Mapping

- Thomas Hunt Morgan's lab had recombination fractions for dozens of genes in *Drosophila*
 - Alfred Sturtevant (19-year-old undergrad) developed the world's first genetic map after skipping one night's homework
 - Four linkage groups in *Drosophila* – four chromosomes



Higher resolution maps

- Phenotype producing genes are few and far between in many eukaryotes (~20k genes total, only a relative few have observable phenotypes)
- More detailed maps rely on DNA sequence variants
 - Restriction digest maps
 - STRs (DNA fingerprint loci)
 - AFLP
 - RAPDs
 - + dozens



Genome Sequencing and Assembly

- **The ultimate map is knowing the position and identity of every nucleotide in a genome, a genome assembly.**
- Two steps:
 - Sequence the DNA – now relatively cheap and easy
 - Assemble the fragments – still kinda hard but getting easier
- How do we do this and how much does it cost?

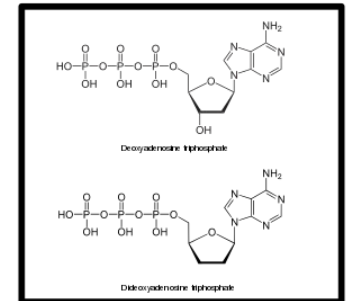
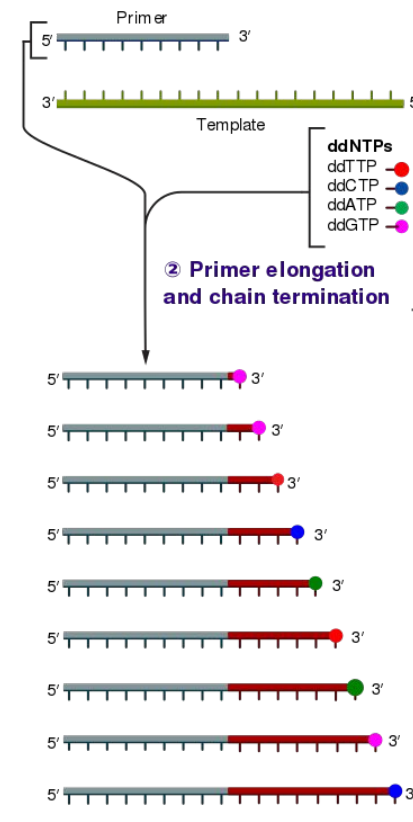
TATTGAATTTTCAAAAATCCTTACTTTTTTTTTTTGGATGGACGCCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATZ
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAACCTTCTCTTTGGAACCTTTG
AATACGCTTAACGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGGGCAGACGCCCTCCGACCGAAGACTCTCCTC
GCGTCTCCTTCTTTCACCGGTGCGCTTCCTGAAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACTI
TCTTTCTGCTTTCAAGTGGTAACTGAGTTGAGTAAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATC
ATCCATTAGTPTTATAGCCITATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAAZ
TCTGAAAACGACTTCTACTAATATCTTCAACATTTTCAGTTTGTATTACTTCTTATTTCAAATGTCATAAAAAGTATCAACAAAAAAT
TCTGAAAACGACTTCTACTAATATCTTCAACATTTTCAGTTTGTATTACTTCTTATTTCAAATGTCATAAAAAGTATCAACAAAAAAT
TACGTC AAGGAGAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGAATTGTACCTGAGTTCAAZ
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGC
CTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGGAT
TCTTGGCCCGTCAAAGTTTTGAAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAZ
CGATTTGCCGTTGGACGGTCTTATGTCAACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAAATGTGGTCTCCATGTTZ
ACTCTTTTCTAAAGAACTTGCAACCGAAAGTTTGCAGTCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGTATGACZ
GGCAGTGGATTTGTCTTCTCGGCCGCAATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAZ
CAAGCAAATTTAATGCGTATTACGGTCTGTTGCAGAACATTATGTTGGTGTAAACAATGGCGGTATGGATCAGGCTGCCTCTGTTZ
GTAGCAGATTTCTAGCTTTTGTACTTTGTGACTTTCAACCGCAGTTGAAGGCTACTCCGTTTTAAATTTCCGCAATTAAAAAACCATGAZ
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTGTTTACTTTTTCTGGAAAAGAAGGATCGAGACCGAATAAAGGTAATCTAAGAZ
TCATGAACGTTTATTATGACCAGATATCACAACATTTCCACACCCCTGGAACGGCGATTTGAAATCCGGCATCGAACGGTTAAACAAC
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTGACGATGTGCGACAATCCTTGAATTTGTTCTCGCZ
ATTTAGCAGAGATTTACTTAAACACTTCCAGTGAAGTTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAZ
TAAAGCTGAAATTTGTAAGTAAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTTGATC
GAGTCTCAAGCTTCTTGGGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTTGCTTTGTCAAATGGATC
TGGTTCCCGTTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAATGGCAACATAGAAAAGGTAZ
AAGCCCTTGCCAATGAGTTCTACAAGTCAAGTACCCTAAGATCATCTACTGATGTGACTGAGGTTAGAAAATGCTATCAGTCTCTAAACZ
TTGGGCAGCGTCTTATGAATTGAAGTATACTTCTTTTTTTTTACTTTGTTTTCAGAACAACTTCTCATTTTTTTTTCTACTCATAACZ
TCAAGTAAAAATCCACTTAAATGAGTAGTAACACTTTTATAGTTTACATCATGCTTTCAACTACTTAAATAAATGATTGTATGATAZ
TTTTCAAGTGAAGAGATTTCCGATTTACCAAAAATTTAAAACACAGGGACAAAATCTTTGATATGCTTTCAACCGCTGCGTTTTGZ
CCTATTCTTGACATGATATGACTACCATTTTTGTTATTGTACGTGGGGCAGTTGACGTTTATCATATGTCAAAGTCAATTTGCGAAZ
TTGGCAAGTTGCCAAC TGACGAGATGCAGTAAAAGAGATTGCGCTTTGAAACTTTTTGTCTCTTTTTTTTTTTCCGGGACTTCAZ
AACCCTTTGTCTACTGATTAATTTTGTACTGAAATTTGGACAATTCAGATTTTAGTAGACAGCGCGAGGAGAAAAGAACTGACZ
AAATTTCCGATGGACAAGAAGATAGGAAAAAAGTCTTACCAGATTTCTTAGACCGAAAAAAGTCTGTATGACATCAGAATGAZ
ATTTTTCAAGTTAGACAAGGACAAAATCAGGACAAAATTTGTAAGATATAATAAACTATTTGATTCAAGCGCAATTTGCCCTTTTCCAZ
TCCATTAATCTCTGTTCTCTCTTACTTATATGATGATTAGGATATCATCTGTATAAAAACCTCCTTCTTAATTTCACTCTAAAGCAZ
CCATAGAGAAGATCTTTCGGTTTCAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATZ
GCGGCTCTTCAAAAAGATTGAACTCTCGCAACTTATGGAACTTTCCAATGAGACCTTTGCGCCAAATAATGTTGGATTTGGAAATZ
TATAAGTCATCTCAGAGTAATAATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAAGTCAGAAAACCTCAAZ
CTCATTCTGGAAGAAAAATCTATTATGAATATGTGGTCGTTGACAAAATCAATCTTGGGTGTTTTCTATTCTGGATTCAATTTATGTACZ
AGGACTTGAAGCCCGTCGAAAAAGAAAGCGGGTTTTGGTCTGGTACAATTTATTGTTACTTCTGGCTTGTGAAATGTTTTCAATATZ
ACTTGGCAAATTTGCAGCTACAGGTCTACAACCTGGGCTCAAAATTTGGTGGCAGTGTGGATAACAATTTGGATTGGGTACGTTTTGCZ
TCCCTTTTTCTTTTCCCTTTCACTTTGCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTT

DNA sequencing methodologies: ca. 1977 AKA 1st generation sequencing

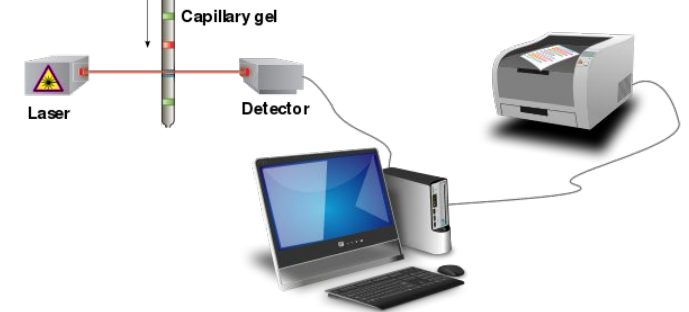
- Sanger/Chain termination sequencing
 - DNA replication based
 - Substitution of substrate with chain-terminator chemical.
 - Efficient
 - Amenable to automation

① Reaction mixture

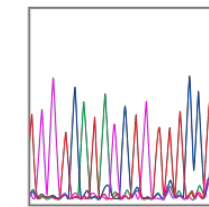
- ▶ Primer and DNA template
- ▶ DNA polymerase
- ▶ ddNTPs with flourochromes
- ▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flourochromes and computational sequence analysis



DNA Analysis: DNA Sequencing

- Best case scenario for Sanger sequencing
 - 700-1000 bp reads
 - 96 capillary machine
 - 3 hour run time
 - ~700 bp/capillary
 - ~5.4 Mb/day
 - ~600 days to sequence the human genome to 1X depth



DNA Analysis: DNA Sequencing

- What is sequencing depth/coverage?
 - Depth is the number of times a given nucleotide has been sequenced, on average
 - ‘Coverage’ is often used interchangeably with ‘depth’
 - Typically, the higher the number, the better the assembly
- 1X coverage = sequencing enough fragments to generate the equivalent of one genome’s worth of sequence data.
 - For the human genome, that would be ~3 billion bases of sequence data
- Is 1X enough to get an accurate genome assembly?

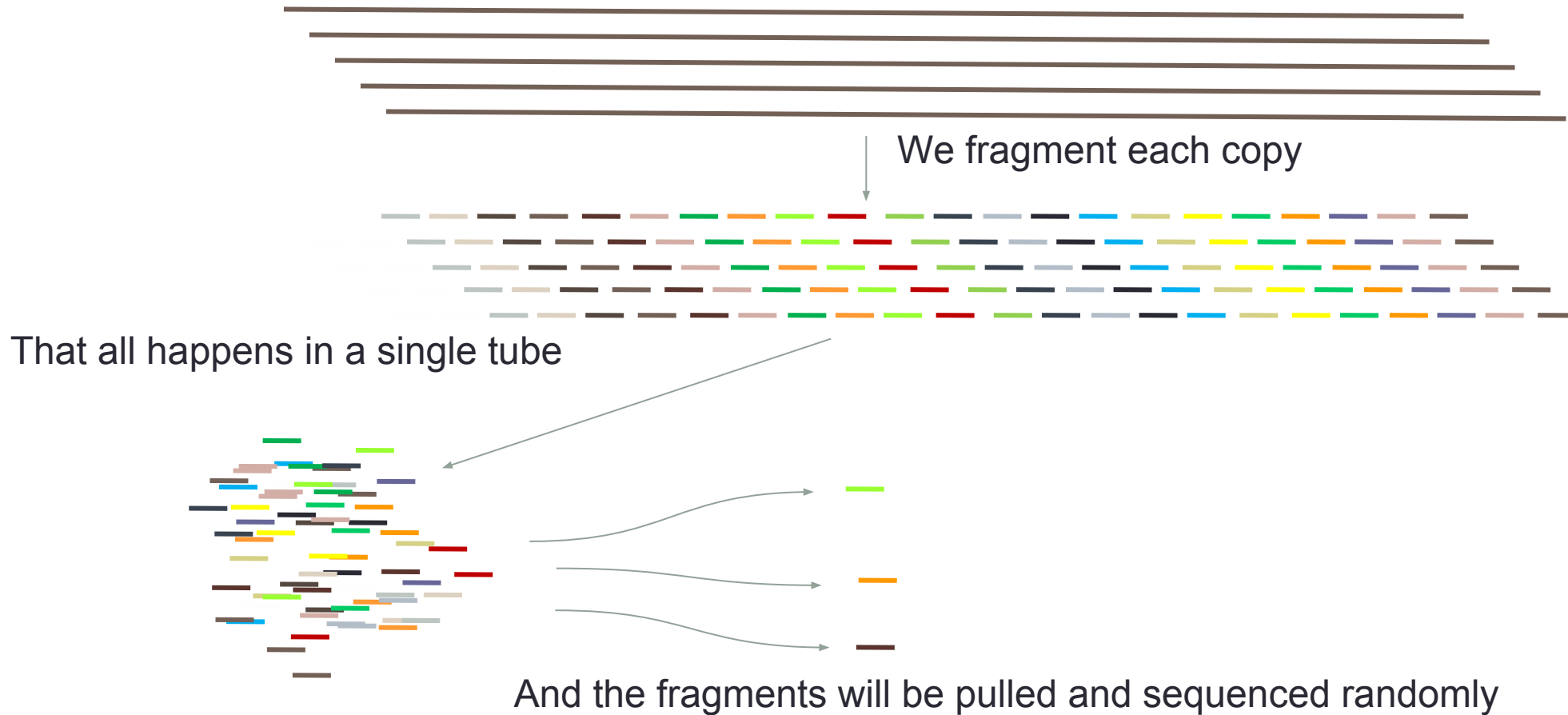
Is 1X enough?

- **No.** Why?

- No current sequencing technologies allow for telomere-to-telomere sequence reads
- Sequence reads that are relatively small compared to chromosome length. We stitch those randomly selected reads together to assemble them
- We also almost always rely on making our sequencing libraries from multiple copies of the species' genome
- Random selection of fragments to sequence
- All sequencing technologies involve error

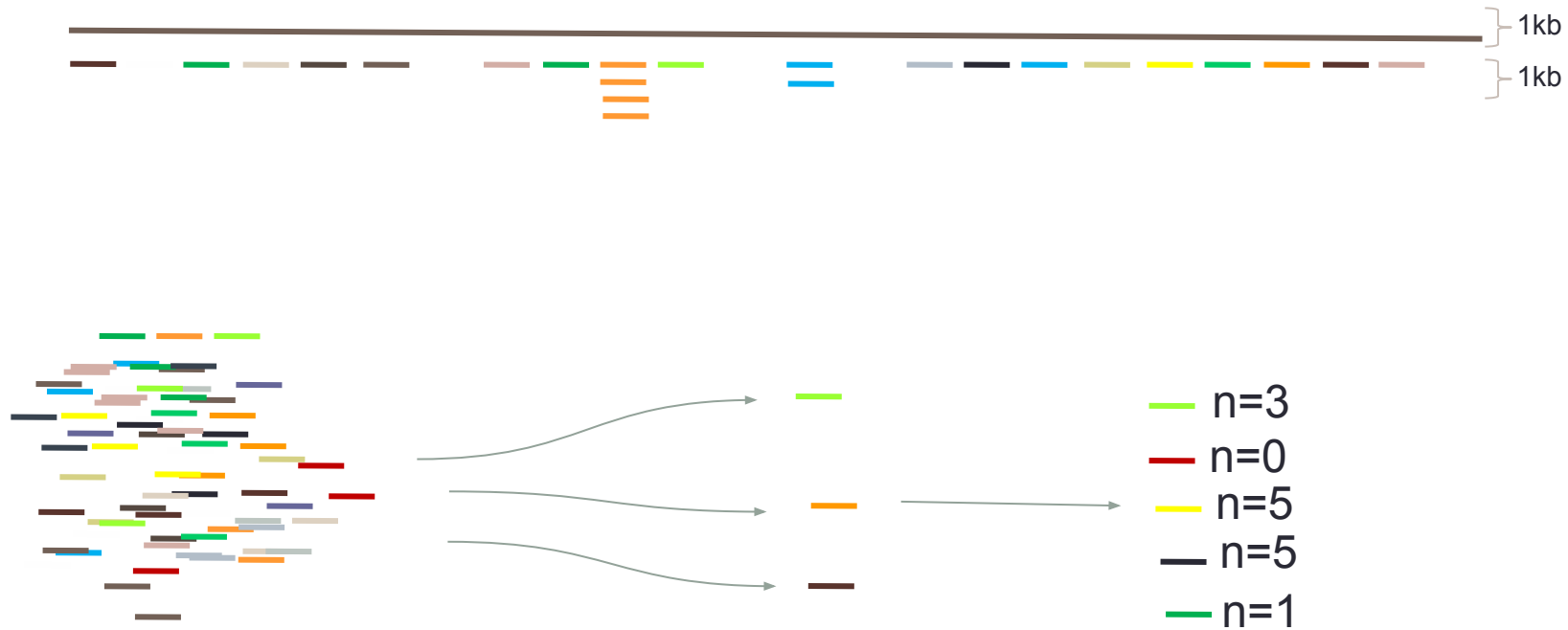
Is 1X enough?

- Imagine you have a genome consisting of only one chromosome and have isolated five cells worth of DNA from that organism.



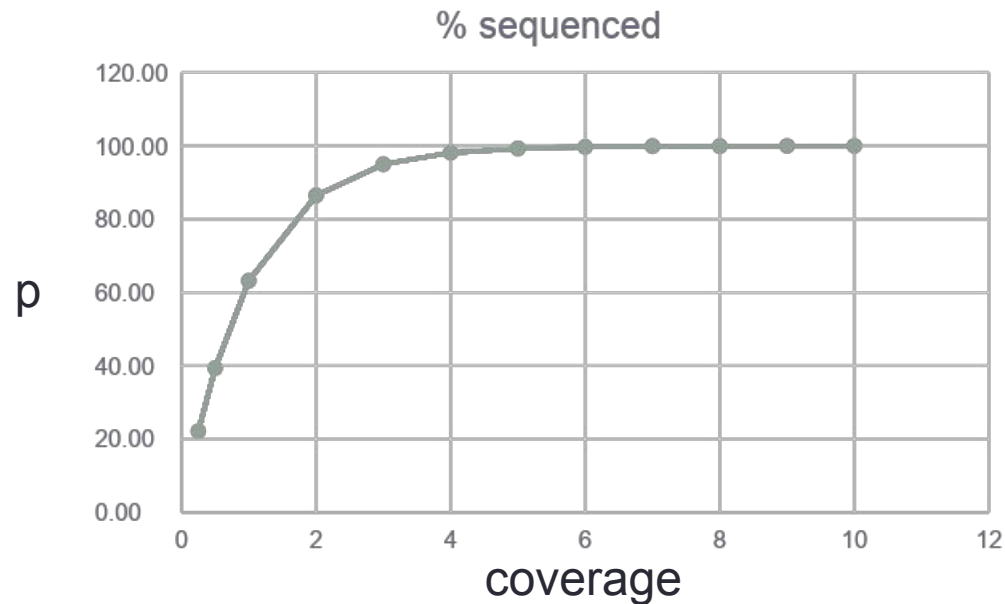
Is 1X enough?

- We'll get some fragments >1 time, some a bunch of times, and some not at all.



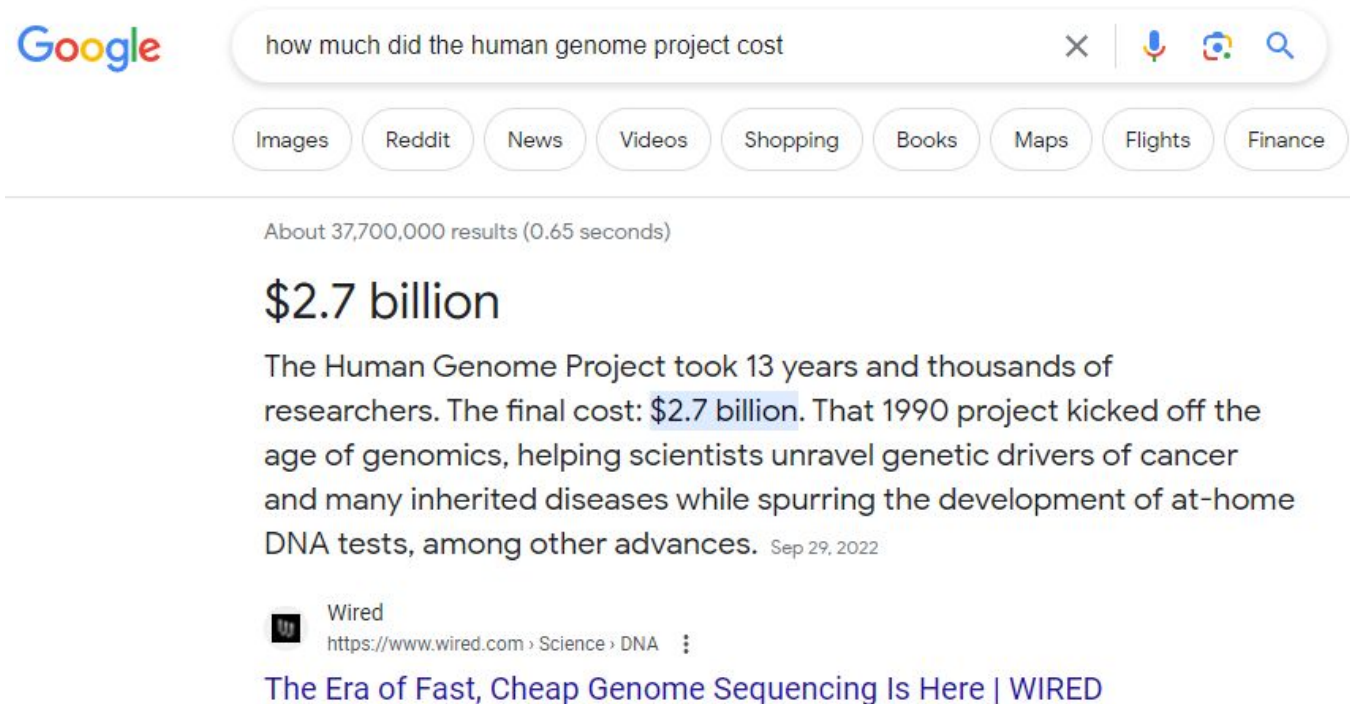
Is 1X enough?

- Probability of a base being sequenced (p) = $1 - e^{-c}$
- C = Coverage
- At 1x, $p = 1 - e^{-1} = 63.21$
- We'll miss ~36.79 % of the genome on average



Genome Assembly

- ▣ **Cost/Mb (1990s, Sanger sequencing) = ~\$9000**
- ▣ **Human genome = 3000Mb**
- ▣ **Cost/1X human genome = \$27,000,000**
- ▣ **Cost/10X human genome = \$270,000,000**
- ▣ **But that's just the sequencing**



The image shows a Google search interface. The search bar contains the text "how much did the human genome project cost". Below the search bar are navigation buttons for Images, Reddit, News, Videos, Shopping, Books, Maps, Flights, and Finance. The search results show "About 37,700,000 results (0.65 seconds)". The main result is a large heading "\$2.7 billion" followed by a paragraph: "The Human Genome Project took 13 years and thousands of researchers. The final cost: \$2.7 billion. That 1990 project kicked off the age of genomics, helping scientists unravel genetic drivers of cancer and many inherited diseases while spurring the development of at-home DNA tests, among other advances. Sep 29, 2022". Below this is a snippet from Wired with the URL "https://www.wired.com › Science › DNA" and the title "The Era of Fast, Cheap Genome Sequencing Is Here | WIRED".

Google

how much did the human genome project cost

Images Reddit News Videos Shopping Books Maps Flights Finance

About 37,700,000 results (0.65 seconds)

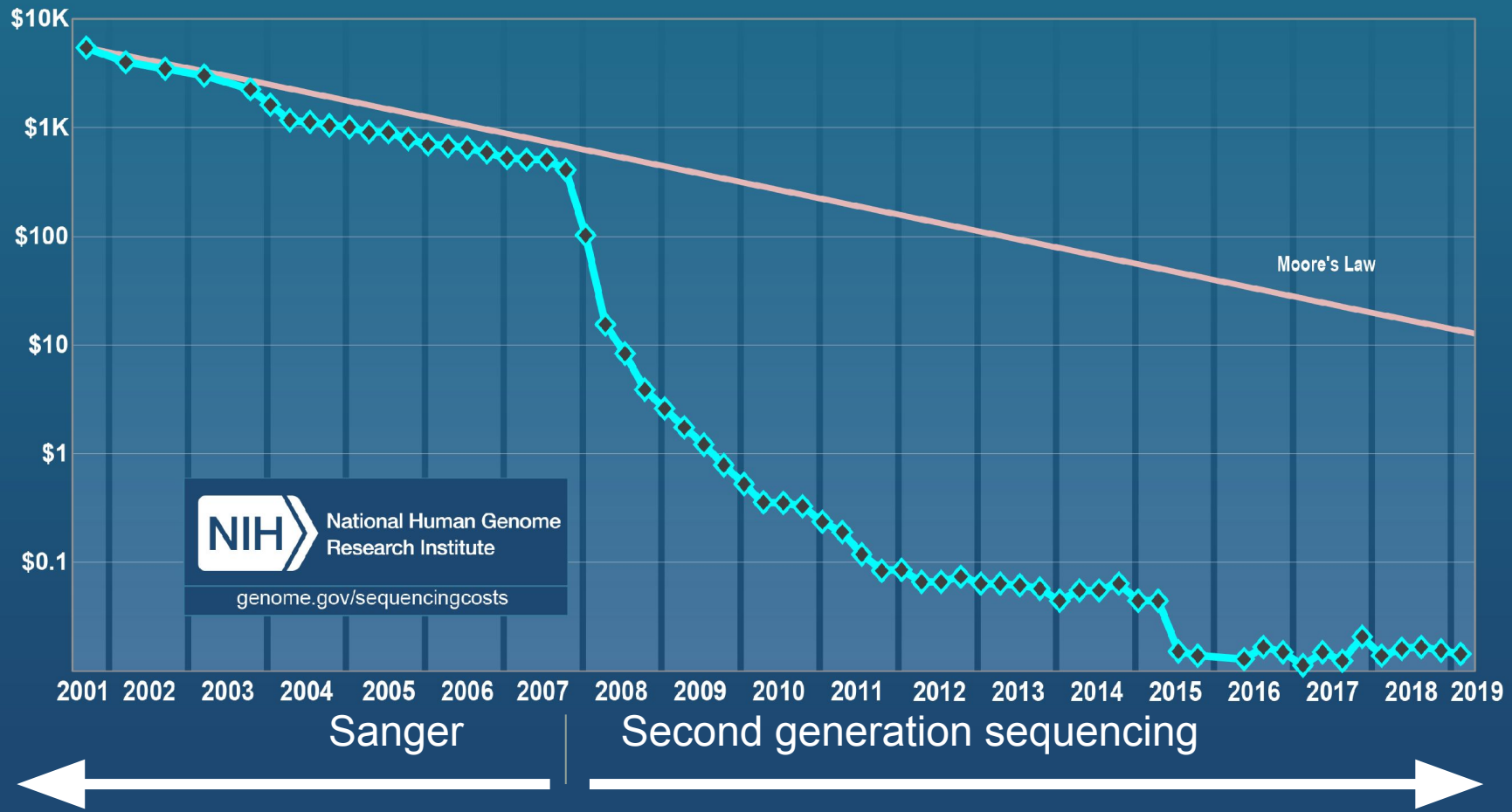
\$2.7 billion

The Human Genome Project took 13 years and thousands of researchers. The final cost: \$2.7 billion. That 1990 project kicked off the age of genomics, helping scientists unravel genetic drivers of cancer and many inherited diseases while spurring the development of at-home DNA tests, among other advances. Sep 29, 2022

Wired
https://www.wired.com › Science › DNA

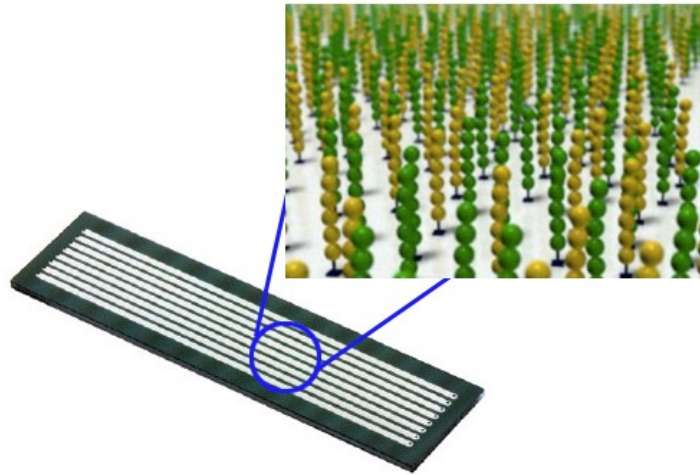
The Era of Fast, Cheap Genome Sequencing Is Here | WIRED

Cost per Raw Megabase of DNA Sequence



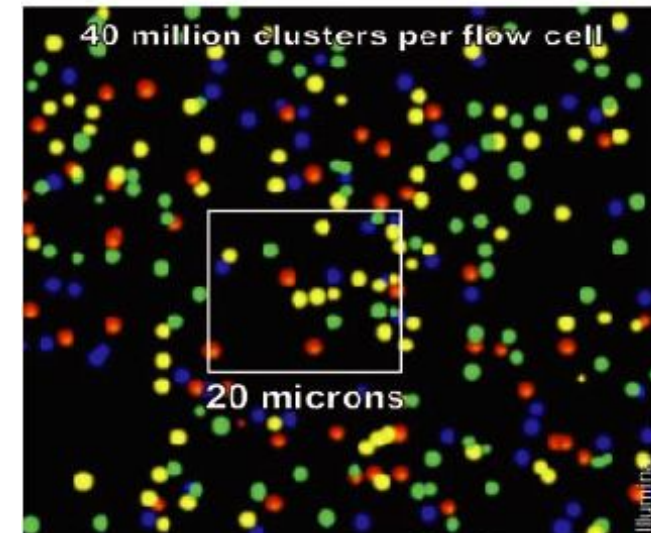
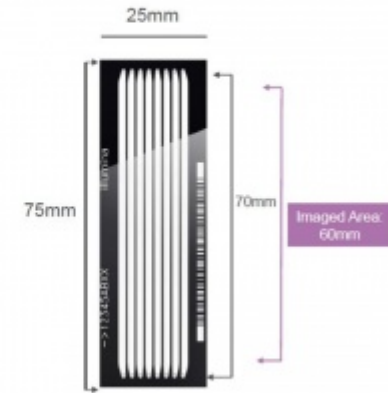
Next (2nd) Generation Sequencing Illumina

- **Illumina** is a massively-parallel sequencing-by-synthesis system
- Really took off ~2007
 - Currently dominates 2nd gen
- Works using flow cells



Flowcell

- 8 channels
- External dimensions: 75 mm x 25 mm
- Channel width: 1.7 mm
- Wall thickness: ~0.4 mm
- Channel depth: 70 μ m
- Ports are on the bottom
- Channel length: 70mm
- Channel imaged area: 60mm



Must see video - <https://www.youtube.com/watch?v=fCd6B5HRaZ8>

Next (2nd) Generation Sequencing

Illumina

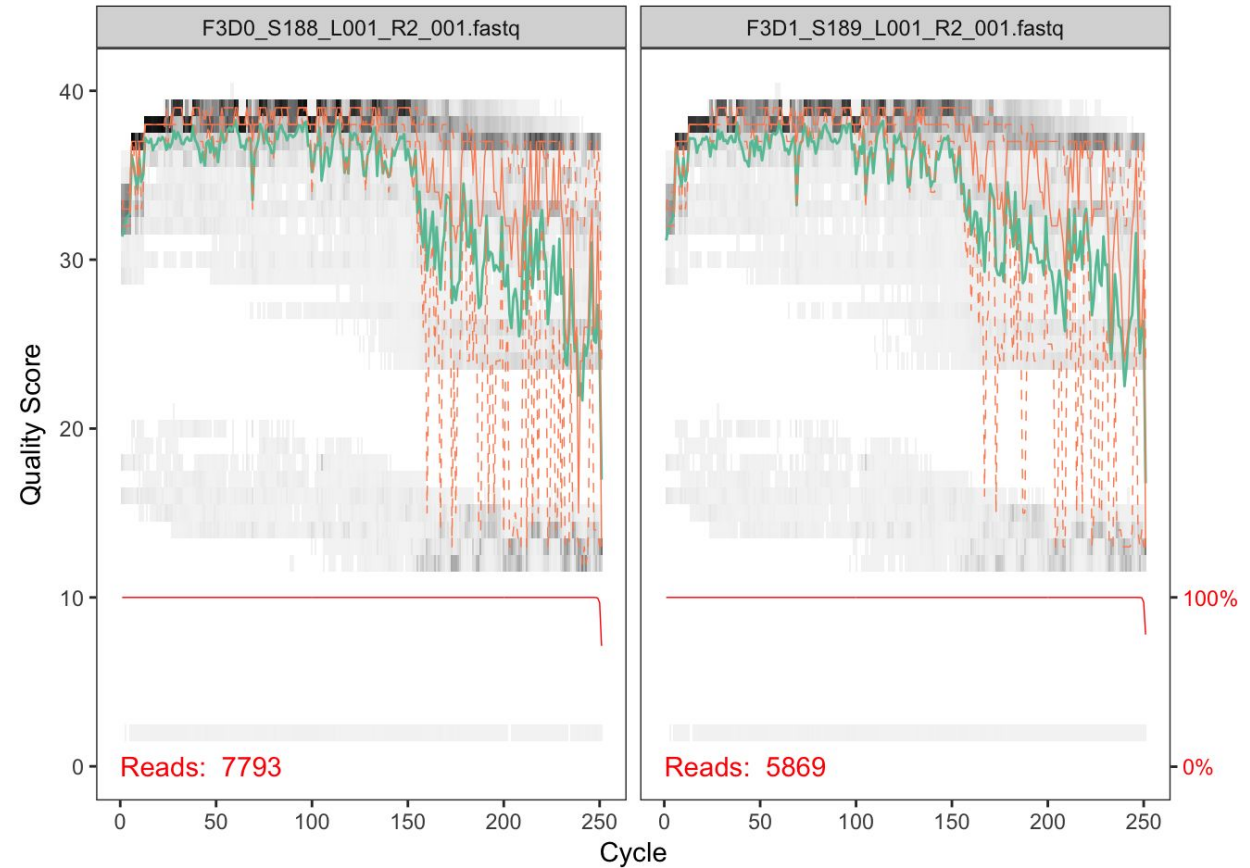
- Pros

- Illumina sequences are highly accurate (99.9+%)
- Cheap
- Library prep is reasonably easy

- Cons

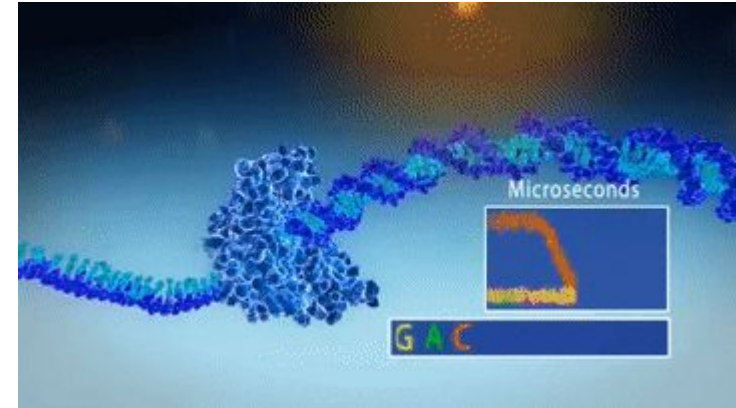
- Maximum read length is ~300 bp

- ▣ **Cost/Gb = \$4-\$200**
- ▣ **Human genome = 3Gb**
- ▣ **Cost/1X human genome = \$12**
- ▣ **Cost/100X human genome = \$120**

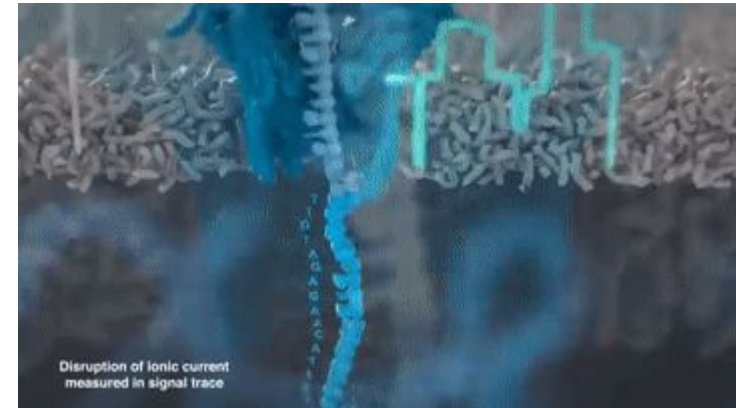


Next (3rd) Generation Sequencing PacBio and Nanopore

- Read lengths 10-40 kb on average
- Performance
 - Long reads – 10 - 40kb on average
 - <1 day run time



<https://www.youtube.com/watch?v=v8p4ph2MAvI>



<https://www.youtube.com/watch?v=RcP85JHLmnl>

Next (3rd) Generation Sequencing PacBio and Nanopore

- Pros

- Cheap but not as cheap
- Library prep is reasonably easy
- Long reads

- Cons

- More expensive
- Higher error rate

- ▣ **Cost/Gb = \$30**

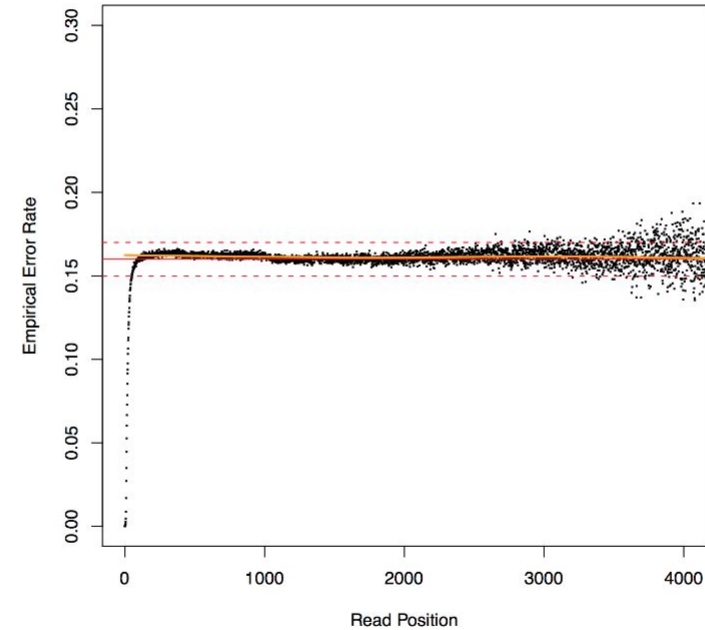
- ▣ **Human genome = 3Gb**

- ▣ **Cost/1X human genome = \$90**

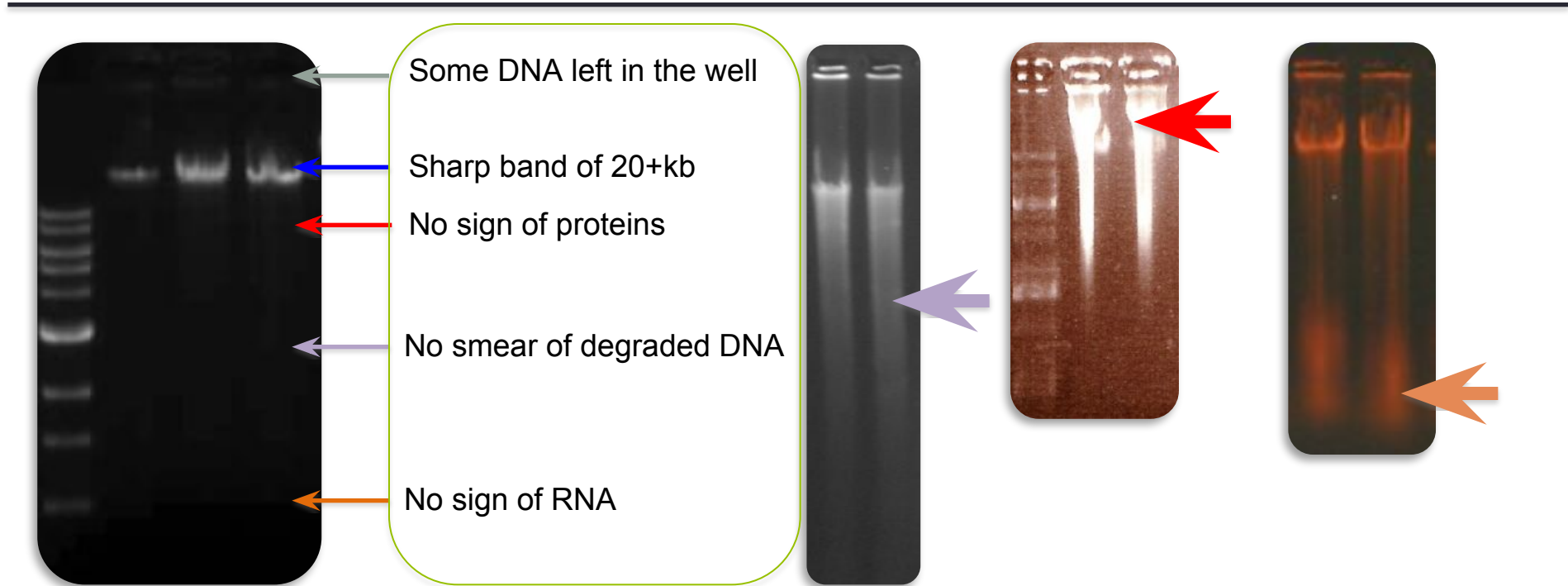
- ▣ **Cost/100X human genome = \$900**

- Error profile

- ~15% as of 2012
- ~13% as of 2017



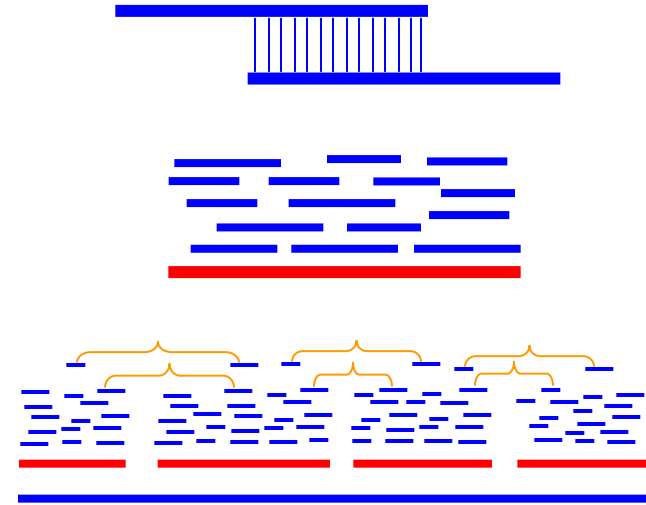
Long read technologies require pure, high molecular weight DNA



- This requires special handling of the tissues prior to DNA extraction
- This can be *logistically difficult* and *cost-prohibitive*.

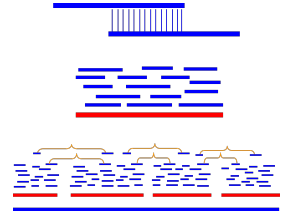
Genome Assembly

- ▣ **Reminders**
 - Getting the sequence is just the first step
 - Reads (even long reads) are very short compared to the length of an entire genome
- ▣ **De novo genome assembly**
 - Obtain reads
 - Assemble reads into contigs
 - Assemble contigs into scaffolds
 - Refine scaffolds into a consensus sequence
 - If possible, map scaffolds to chromosomes



..ACGATTACAATAGGTT..

An assembly problem



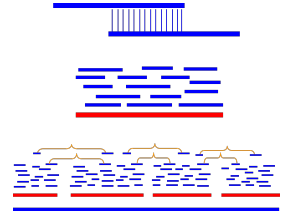
Single-end reads

```
tttggtcgca    gtcgcccta
tccagcatca    cggctagata
ccagcaccac    cacttagccg
ggctagatat    tttggtcgcc
tcacttagcc    ccctatacg
gccgaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
tttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



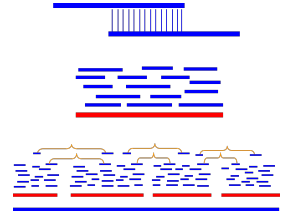
Single-end reads

```
tttggtcgca  
tccagcatca  
ccagcaccac  
ggctagatat  
tcacttagcc  
gccgaaaaag  
gtcgcccta  
cggctagata  
cacttagccg  
tttggtcgcc  
ccctatagc
```

Paired-end reads

```
tgggtgt-----cgaaaaa  
ggtagcc-----attgggt  
tttggtc-----gatattg  
agatatt-----cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc
```

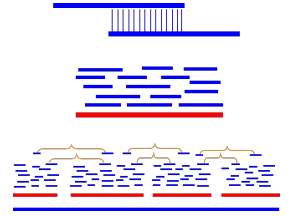
Single-end reads

```
tccagcatca
ccagcaccac
ggctagatat
tcacttagcc
gccgaaaaag
gtcgcccta
cggctagata
cacttagccg
cccctatag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
tttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
  ccctatacg
```

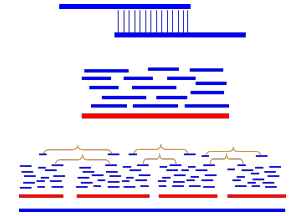
Single-end reads

```
tccagcatca    cggctagata
ccagcaccac    cacttagccg
ggctagatat
tcacttagcc
gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
tttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgcccta

ccctatacg

Single-end reads

tccagcatca

ccagcaccac

ggctagatat

tcacttagcc

gccgaaaaag

cggctagata

cacttagccg

Paired-end reads

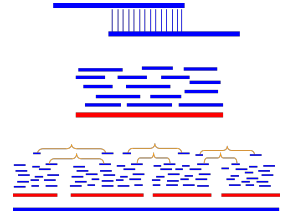
tgggtgt-----cgaaaaa

ggtagcc-----attgggt

tttggtc-----gatattg

agatatt-----cttagct

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
  ccctatacg

tcacttagcc
  cacttagccg

cggctagata
  ggctagatat
```

Single-end reads

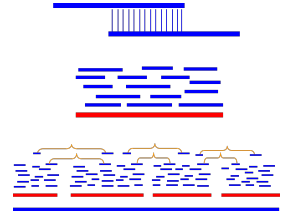
```
tccagcatca
ccagcaccac

gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```


An assembly problem



```
t ttggtcgca
t ttggtcgcc

g tcgccccta
  ccctatacg

t cacttagcc
  cacttagccg

c ggctagata
  ggctagatat

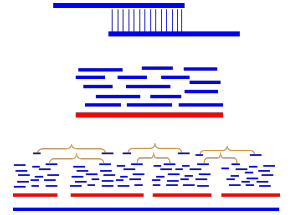
t ccagcatca
  ccagcaccac

g ccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
gtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
t ttggtcgca
t ttggtcgcc

gtcgcccta
  ccctatacg
```

```
tcacttagcc
  cacttagccg
```

```
cggctagata
  ggctagatat
```

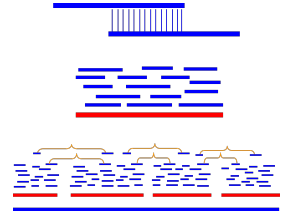
```
tccagcatca
  ccagcaccac
```

```
gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
t ttggtcgca
t ttggtcgcc

g tcgccccta
  ccctatacg
```

```
c ggctagata
  ggctagatat
```

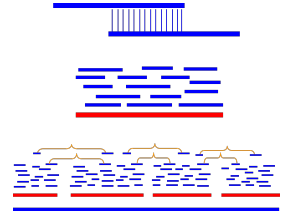
```
t ccagcatca
  ccagcaccac
```

```
tcacttagcc
 cacttagccg
  gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
tttggtcgca
tttggtcgcc

gtcgccccta
  ccctatacg
```

```
cggctagata
  ggctagatat
```

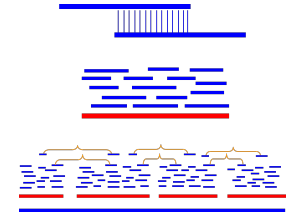
```
tccagcatca
  ccagcaccac
```

```
tcacttagcc
  cacttagccg
    gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



tttggtcgca

tttggtcgcc

gtcgccccta

ccctatagc

cggtagata

ggctagatat

tccagcatca

ccagcaccac

tcacttagcc

cacttagccg

gccgaaaaag

Paired-end reads

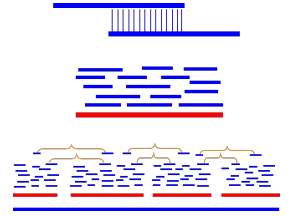
tgggtgt-----cgaaaaa

ggtagcc-----attgggt

tttggtc-----gatattg

agatatt-----cttagct

An assembly problem



```
gtcgccccta
  ccctatagc
```

```
tttggtcgca
tttggtcgcc
```

```
cggctagata
  ggctagatat
```

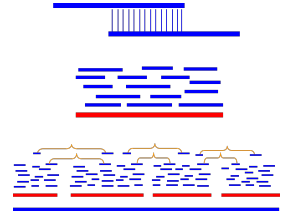
```
tccagcatca
  ccagcaccac
```

```
tcacttagcc
  cacttagccg
    gccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



gtcgcccta
ccctatag

ttggtcgca
ttggtcgcc

tcacttagcc
cacttagccg
gccgaaaaag

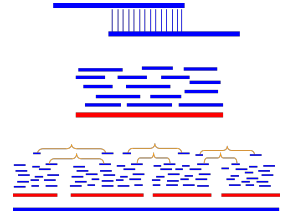
cgctagata
ggctagatat

tccagcatca
ccagcaccac

Paired-end reads

tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct

An assembly problem



```
tttggtcgca
tttggtcgcc
    gtcgcccta
        ccctatacg
```

```
tcacttagcc
cacttagccg
    gccgaaaaag
```

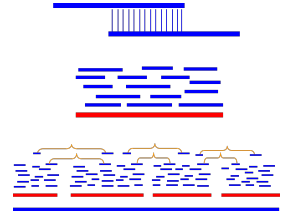
```
cggctagata
  ggctagatat
```

```
tccagcatca
  ccagcaccac
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```


An assembly problem



```
tttggtcgca
tttggtcgcc
    gtcgcccta
        ccctatacg
```

```
tcacttagcc
cacttagccg
    gccgaaaaag
```

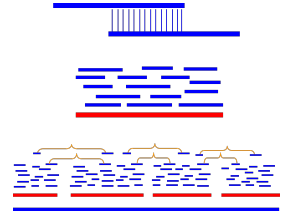
```
cggctagata
ggctagatat
```

```
tccagcatca
ccagcaccac
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
cgctagata
ggctagatat
```

```
tttggtcgca
tttggtcgcc
    gtcgcccta
        ccctatacg
```

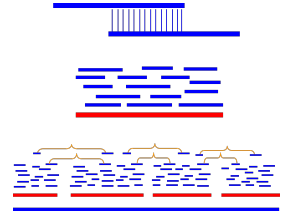
```
tcacttagcc
cacttagccg
    gccgaaaaag
```

```
tccagcatca
ccagcaccac
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem



```
cgctagata
ggctagatat
cgctagatat
```

```
tttggtcgca
tttggtcgcc
    gtcgcccta
        ccctatacg

tttggtcgccctatacg
```

```
tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac
tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```

An assembly problem

```
tttggtcgca
tttggtcgcc
  gtcgcccta
    ccctatacg

cggctagata
  ggctagatat

tcacttagcc
  cacttagccg
tccagcatca  gccgaaaaag
  ccagcaccac

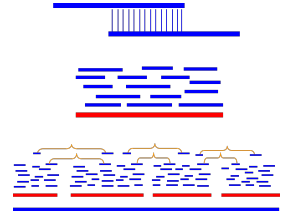
cggctagatat
```

```
tttggtcgccctatacg
```

```
tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
ttggtc-----gatattg
agatatt-----cttagct
```



An assembly problem

```
tttggtegca
tttggtegcc
  gtcgccccta
    ccctatacg

cggctagata
  ggctagatat

tcacttagcc
  cacttagccg
tccagcatca  gccgaaaaag
  ccagcaccac

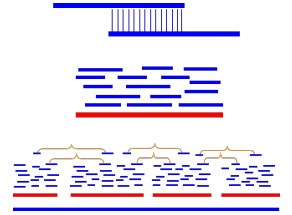
cggctagatat
```

```
tttggtegcccctatacg
```

```
tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgt-----cgaaaaa
ggtagcc-----attgggt
tttggtc-----gatattg
agatatt-----cttagct
```



An assembly problem

tttggtegca
tttggtegcc
gtcgccccta
ccctatacgc

cggtagata
ggctagatat

tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac

cggtagatat

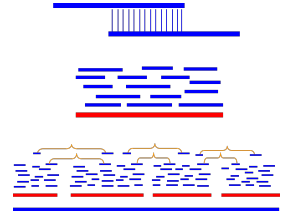
tttggtegcccctatacgc
tttggtc-----gatattg

tccagcatcacttagccgaaaaag

Paired-end reads

tgggtg-----cgaaaaa
ggtagcc-----attgggt

agatatt-----cttagct



An assembly problem

```
tttggtagca
tttggtagcc
gtcgcccta
ccctatacg

cggtagata
ggctagatat

tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac
```

cggtagat

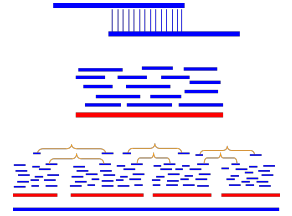
```
tttggtagccctatacg
tttggtagcc-----gatattg
```

tccagcatcacttagccgaaaaag

Paired-end reads

```
tggtagt-----cgaaaaa
gtagcc-----attgggt

agatatt-----cttagct
```



An assembly problem

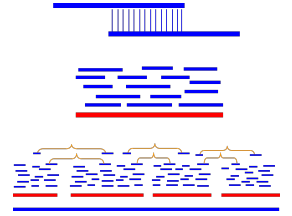
```
tttggtagca      cggctagata      tcacttagcc
tttggtagcc      ggctagatat      cacttagccg
gtcgccccta      tccagcatca      gccgaaaaag
ccctatacag      ccagcaccac
```

```
          cggctagatat
tttggtagcccctatacag
tttggtagc-----gatattg
```

```
tccagcatcacttagccgaaaaag
```

Paired-end reads

```
tgggtgtagc-----cgaaaaa
gtagcatt-----attgggt
agatatt-----cttagct
```



An assembly problem

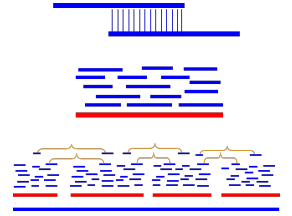
```
tttggtegca      cggctagata      tcacttagcc
tttggtegcc      ggctagatat      cacttagccg
gtcgccccta      tccagcatca      gccgaaaaag
ccctatacag      ccagcaccac
```

```
          cggctagatat
tttggtegcccctatacg
tttggtc-----gatattg
```

tccagcatcacttagccgaaaaag

Paired-end reads

```
tgggtg-----cgaaaaa
ggtagcc-----attgggt
agatatt-----cttagct
```



An assembly problem

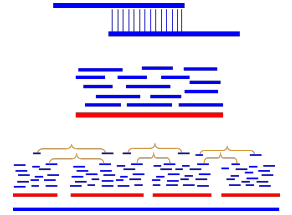
```
tttggtagca      cggctagata      tcacttagcc
tttggtagcc      ggctagatat      cacttagccg
gtcgcccta      tccagcatca      gccgaaaaag
ccctatacg      ccagcaccac
```

```
          cggctagatat
tttggtagccctatacg
tttggtagc-----gatattg
          agatatt-----cttagct
```

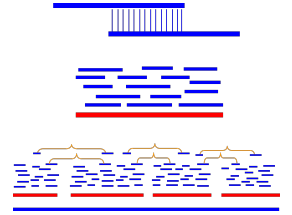
tccagcatcacttagccgaaaaag

Paired-end reads

```
tgggtgtagcc-----cgaaaaa
gtagcc-----attgggt
```



An assembly problem



```
tttggtagca          cggctagata          tcacttagcc
tttggtagcc          ggctagatat          cacttagccg
gtcgccccta          tccagcatca          gccgaaaaag
ccctatacg           ccagcaccac
```

```
          cggctagatat
tttggtagcccctatacg
tttggtc-----gatattg   tccagcatcacttagccgaaaaag
          agatatt-----cttagct
```

Paired-end reads

```
tgggtg-----cgaaaaa
gtagcc-----attgggt
```

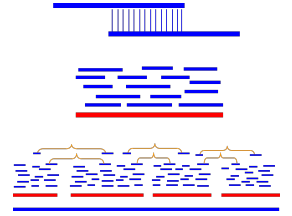
An assembly problem

```
tttggtagca      cggctagata      tcacttagcc
tttggtagcc      ggctagatat      cacttagccg
gtcgccccta      tccagcatca      gccgaaaaag
ccctatacag      ccagcaccac
```

```
          cggctagatat
tttggtagcccctatacag
tttggtagc-----gatattg      tccagcatcacttagccgaaaaag
          agatatt-----cttagct
```

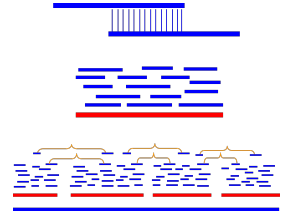
Paired-end reads

```
tttggtagc-----cgaaaaa
gtagcc-----attgggt
```



An assembly problem

```
tttggtagca      cggctagata      tcacttagcc
tttggtagcc      ggctagatat      cacttagccg
gtcgccccta      tccagcatca      gccgaaaaag
ccctatacag      ccagcaccac
```



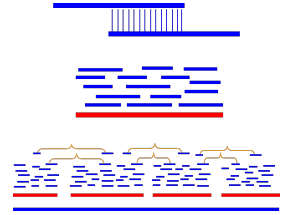
```
          cggctagatat
tttggtagcccctatacg
tttggtagc-----gatattg      tccagcatcacttagccgaaaaag
          agatatt-----cttagct
          tgggtg-----cgaaaaa
```

Paired-end reads

```
gtagcc-----attgggt
```

An assembly problem

```
tttggtagca      cggctagata      tcacttagcc
tttggtagcc      ggctagatat      cacttagccg
gtcgccccta      tccagcatca      gccgaaaaag
ccctatacag      ccagcaccac
```



```
          cggctagatat
tttggtagcccctatac
tttggtagc-----gatattg      tccagcatcacttagccgaaaaag
          agatatt-----cttagct
          tgggtgt-----cgaaaaa
```

Paired-end reads

gtagcc-----attgggt

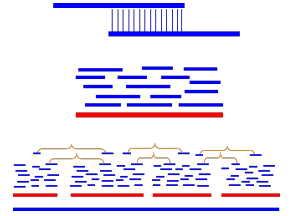
An assembly problem

```
tttggtagca
tttggtagcc
gtcgccccta
ccctatacg

cggctagata
ggctagatat

tcacttagcc
cacttagccg
tccagcatca gccgaaaaag
ccagcaccac
```

```
          cggctagatat
tttggtagcccctatacg
tttggtagc-----gatattg   tccagcatcacttagccgaaaaag
          agatatt-----cttagct
                tgggtgt-----cgaaaaa
ggtagcc-----attgggt
```

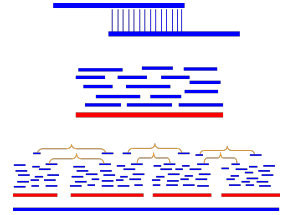


An assembly problem

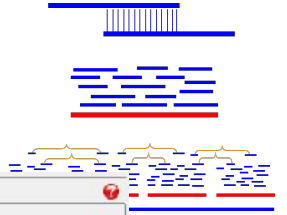
```
tttggtegca      cggctagata      tcacttagcc
tttggtegcc      ggctagatat      cacttagccg
  gtcgcccta      tccagcatca      gccgaaaaag
    ccctatacg      ccagcaccac
```

```
          cggctagatat
tttggtegccccctatacg
tttggtc-----gatattg      tccagcatcacttagccgaaaaag
          agatatt-----cttagct
                tgggtgt-----cgaaaaa
ggtagcc-----attgggt
```

tttggtegccccctatacggctagatatgggtgtccagcatcacttagccgaaaaag



An assembly problem



The screenshot displays a genomic assembly software interface. The top toolbar includes various navigation and visualization options such as 'Open Assembly', 'Import Features', 'Import Contigs', 'Head Loading', 'Tag Variants', 'Reval Colors', 'Zoom', 'Variants', 'Page Left', 'Page Right', 'Jump to Base', 'Prev Feature', 'Next Feature', 'Prev View', 'Next View', 'Head Info', 'Show Bases', 'Reval Names', 'RS Off', 'RS Center', 'RS Column', 'Show Cigar', and 'Sticky Highlights'. Below the toolbar, a table lists contigs:

Contig	Length	Seeds	Features	Mismatch %
1	1,696,055	3,944,293	3	0.4

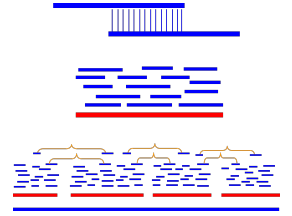
The main area shows a large track of reads with a color-coded alignment. Two specific reads are highlighted with callouts:

- DGL070R1:205.C6WR4ACXX:6:1107:9282:11171**
From: 77 L717 to 617 U317
Length: 101 U101 (0 mismatches)
Cigar: 101M
Properly paired (2/2), insert size: 264
- DGL070R1:205.C6WR4ACXX:6:1107:9282:11171 (Mate)**
From: 300 L830 to 503 U300
Length: 101 U101 (0 mismatches)
Cigar: 101M
Properly paired (1/2), insert size: 264

At the bottom, a 'Tablet Tip' reads: 'Right click on the features track to access the option to select which tracks are visible'. The system clock in the bottom right corner shows 8:20 AM.

An assembly problem

- Sequencing errors

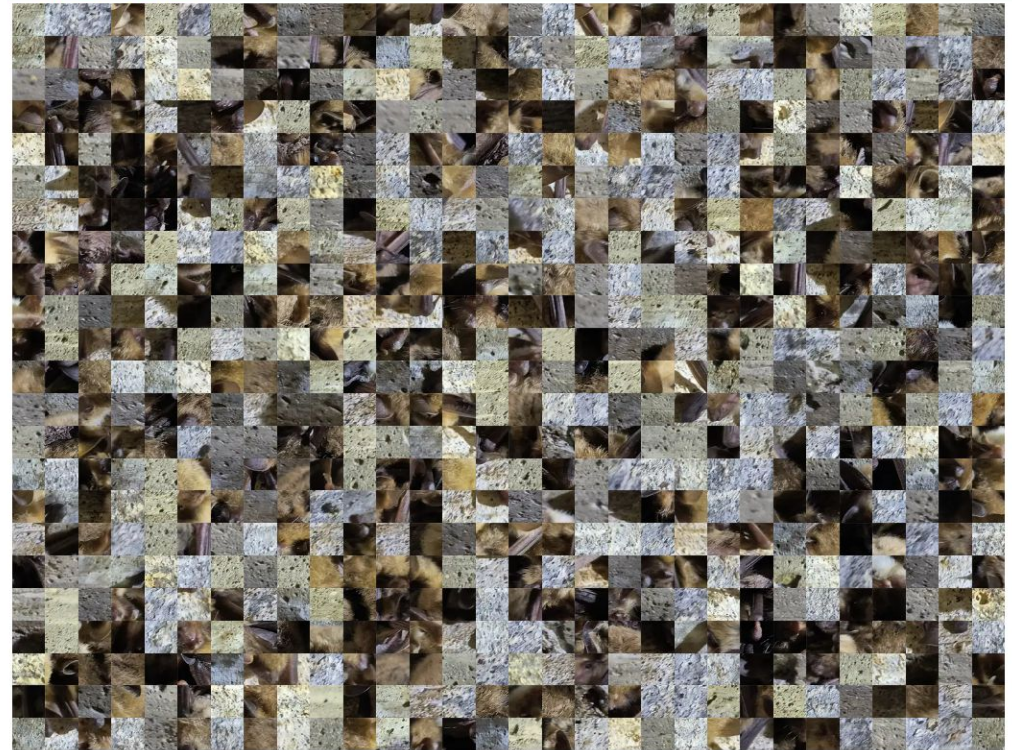


```
tttggtcgda
tttggtcgcc      cggctagata      tcacttagcc
gtcgcccta      ggctagatat      cacttagccg
      ccctatacg      tccagcatca      gccgaaaaag
tttggtc-----gatattg      ccagcacac
      agatatt-----cttagct
      tgggtgt-----cgaaaaa
ggtagcc-----attgggt

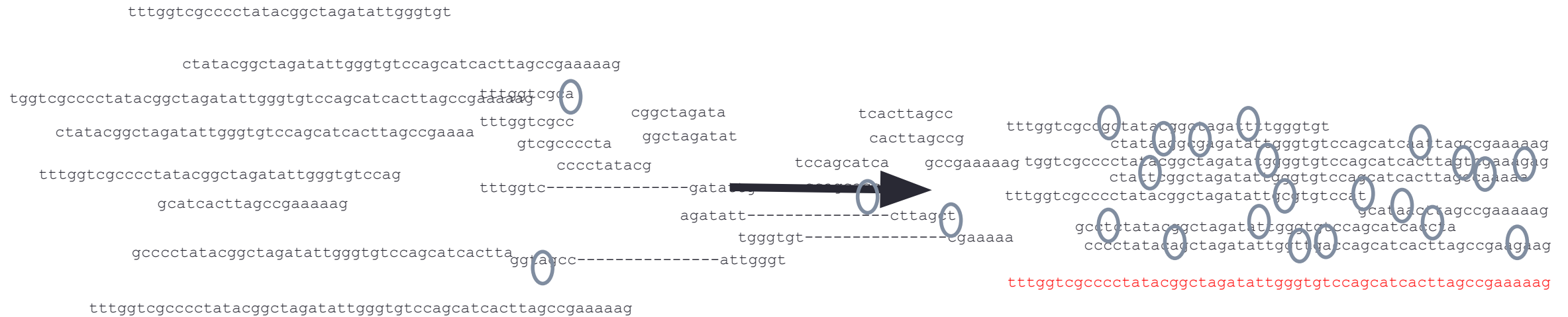
tttggtcgccctatacggctagatattgggtgtccagcatcacttagccgaaaaag
```

How do long reads help with assembly?

- Which puzzle is easier to solve?

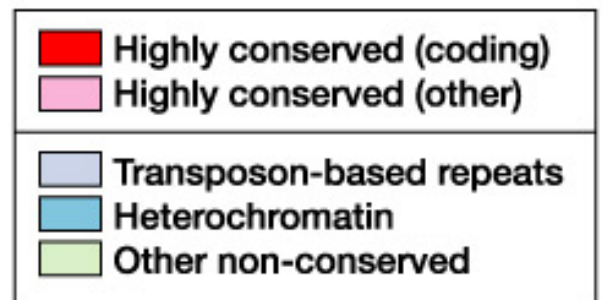
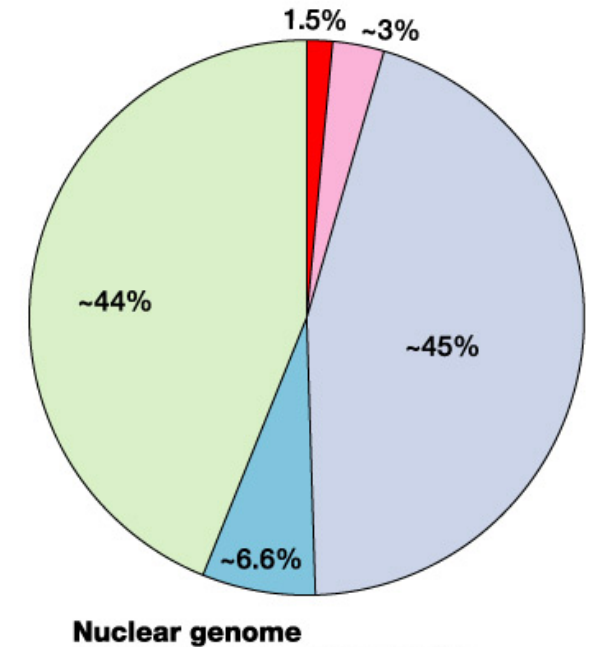
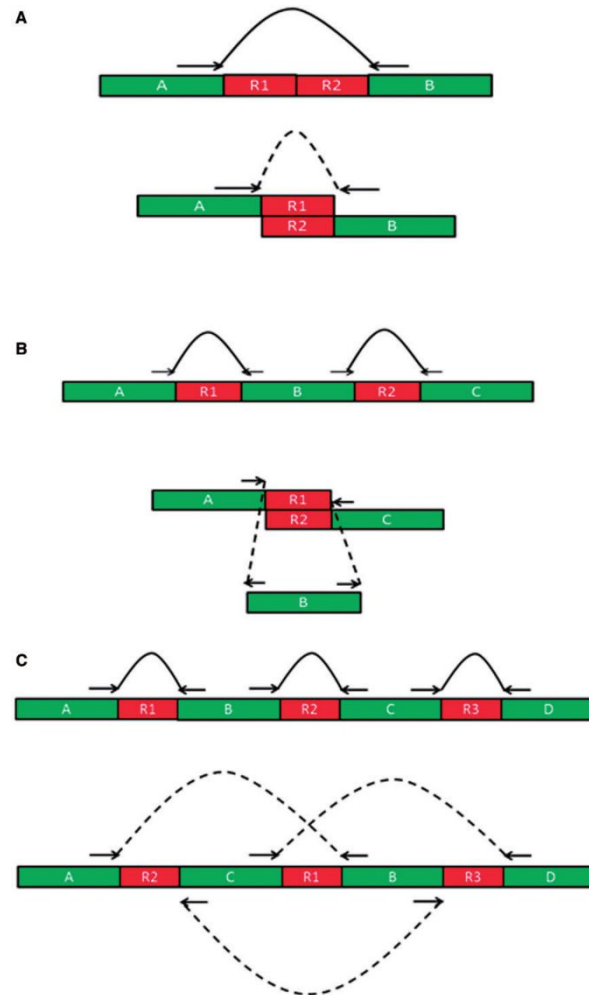


The same assembly problem (but with long reads)



What makes genome assembly difficult?

- Massive scale – billions and billions of reads
- Uneven coverage
- Errors in sequencing reads
- Repetitive regions



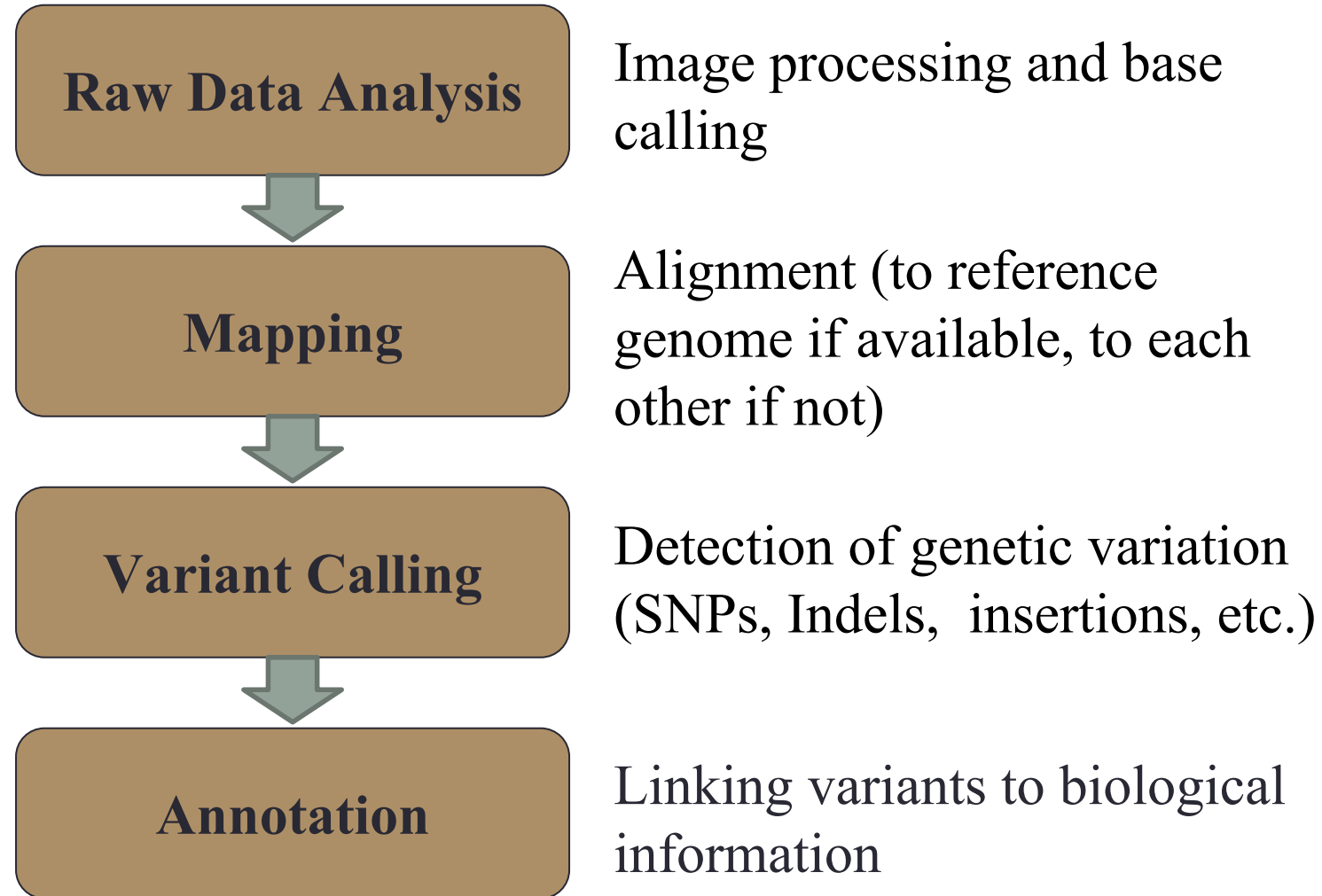
Evaluating an assembly

- How do we know if an assembly is any good?
 - Contiguity – How connected are all the pieces?
 - N50 – Minimum scaffold/contig length at which you've covered 50% of the total assembly length
 - Higher value = better
 - Depth
 - Good numbers depend on the sequencing technology
 - Long reads – 50-60X+, Illumina only – 100X+
 - Fidelity – How well does the raw data match the final product?
 - Map the raw reads back to the assembly
 - Higher value = better
 - Gene content – Are the genes we expect to see present in the assembly?
 - BUSCO – Benchmarking Universal Single-Copy Orthologs
 - Higher value = better

Post-assembly applications

- A genome assembly is great but using it to answer biological questions is the real goal
- “What variations in genomes are important for biology?”
- “How does the genome influence biological diversity?”
- Variation can be:
 - **SNPs – single nucleotide polymorphisms**
 - Insertions
 - Deletions
 - Duplications
 - Inversions – >2 bp sequence reversals at particular loci

Detecting Variation



Read mapping

SNP – heterozygote variant

```
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
  CAGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
    GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
        TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
          TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
            GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
              TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
```

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

reference genome

SNP – homozygote variant

```
ATCCTGATTCGGTGAACGTTATCGACGTTCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGTTCCGATCGA
  CAGTGAACGTTATCGACGTTCCGATCGAACTGTCAGC
    GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
        TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
          TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
            GTTATCGACGTTCCGATCGAACTGTCAGCGGCAAGCT
              TTATCGACGTTCCGATCGAACTGTCAGCGGCAAGCT
```

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

reference genome

Read mapping

- Appropriate read depth is important to identifying variants

SNP – heterozygote variant

```
ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGA
      CAGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
      GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
      TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
```

ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

reference genome

SNP – homozygote variant

```
ATCCTGATTCCGGTGAACGTTATCGACGTTCCGATCGA
ATCCTGATTCCGGTGAACGTTATCGACGTTCCGATCGA
      CAGTGAACGTTATCGACGTTCCGATCGAACTGTCAGC
      GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
      GTTATCGACGTTCCGATCGAACTGTCAGCGGCAAGCT
      TTATCGACGTTCCGATCGAACTGTCAGCGGCAAGCT
```

ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

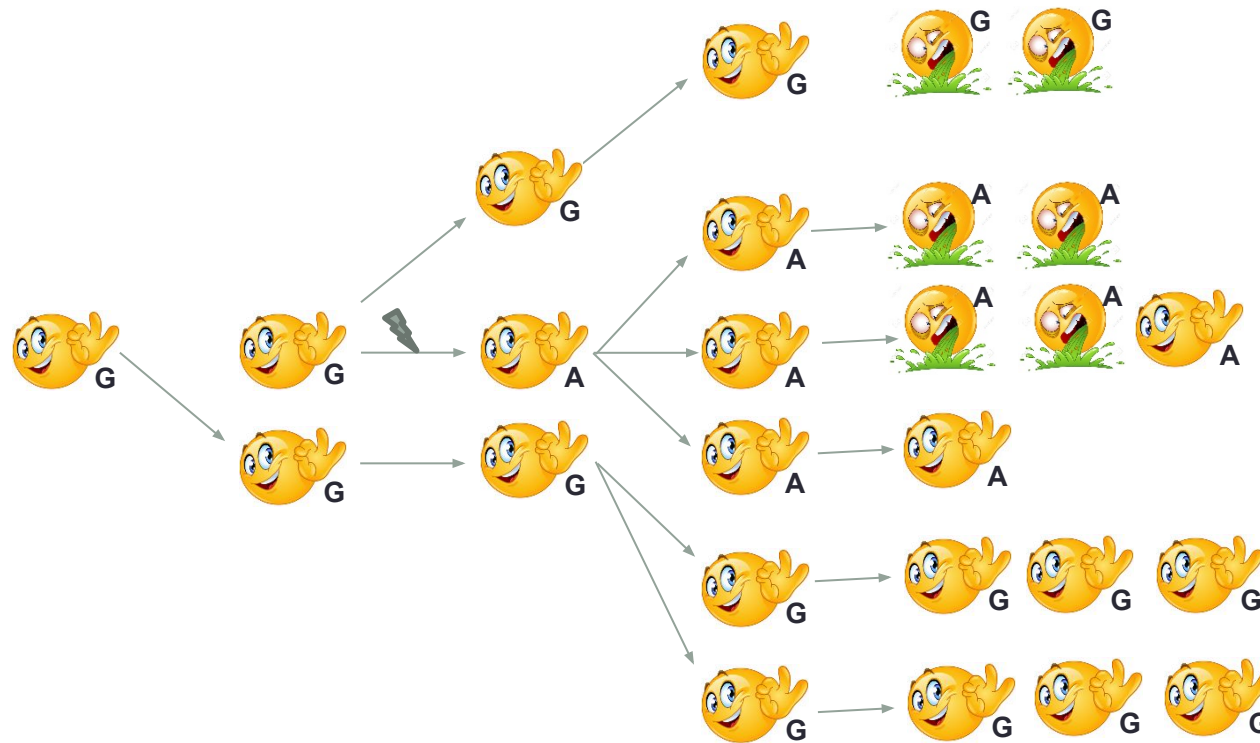
reference genome

What variants and genes are important to phenotype? - GWAS

- ▣ A **genome-wide association study (GWAS)** - identifies SNPs across complete genomes of many samples to find genetic variations statistically associated with a particular phenotype.
- ▣ What do you need?
 - An annotated reference genome (usually)
 - A map of genetic variation
 - A set of technologies that can quickly and accurately analyze whole genomes for genetic variants
 - This is typically accomplished using low coverage sequencing (4-20X).

The basic idea

The A allele is associated (4/14, 29%)
with individuals exhibiting the disease phenotype



GWAS and age-related macular degeneration

- Study cohort – 2172 unrelated individuals of European descent, at least 60 years old
 - 1238 with AMD, 934 controls
 - Each individual harbors two alleles
 - 2476 AMD alleles
 - 1868 non-AMD alleles
- Null hypothesis – Alleles will be randomly distributed in the population, i.e. no association of any alleles with AMD
- Alternative hypothesis – Some allele will be positively associated with AMD

934
controls



1238
cases



Age-related macular degeneration

- Single SNP identified by GWAS, rs1061170
- 4344 alleles recovered, two variants C/T

Allele	Cases with AMD	Controls	Total Alleles
C	1522	670	2192
T	954	1198	2152
Total alleles	2476	1868	4344

- X^2 test suggests association, $p=1.2 \times 10^{-62}$

Evolutionary history - *Myotis* phylogenomics

SINE-Based Phylogenomics Reveal Extensive Introgression and Incomplete Lineage Sorting in *Myotis*

by  Jennifer M. Korstian ¹ ,  Nicole S. Paulat ¹ ,  Roy N. Platt II ² ,  Richard D. Stevens ³  and  David A. Ray ^{1,*}  

¹ Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA

² Host-Pathogen Interactions Program, Texas Biomedical Research Institute, San Antonio, TX 78227, USA

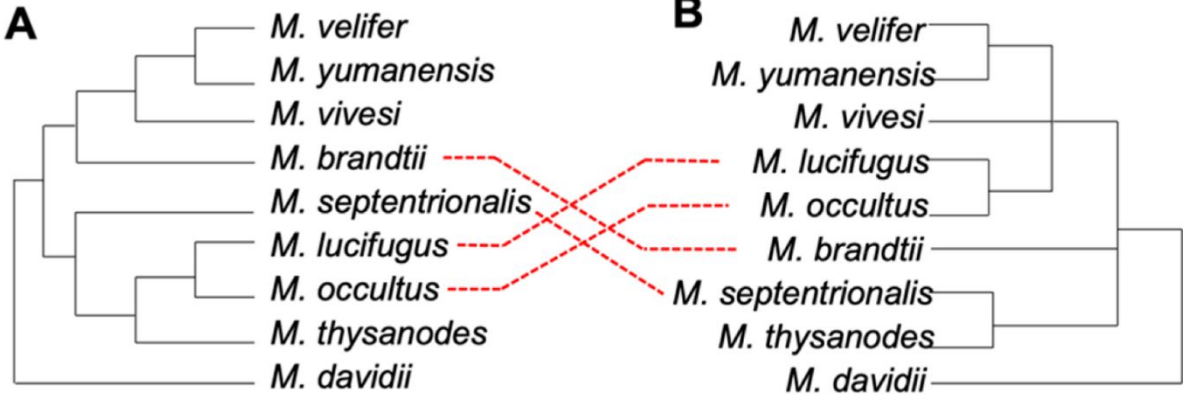
³ Department of Natural Resource Management and Natural Science Research Laboratory of the Museum of Texas Tech, Texas Tech University, Lubbock, TX 79409, USA

* Author to whom correspondence should be addressed.

Genes **2022**, *13*(3), 399; <https://doi.org/10.3390/genes13030399>

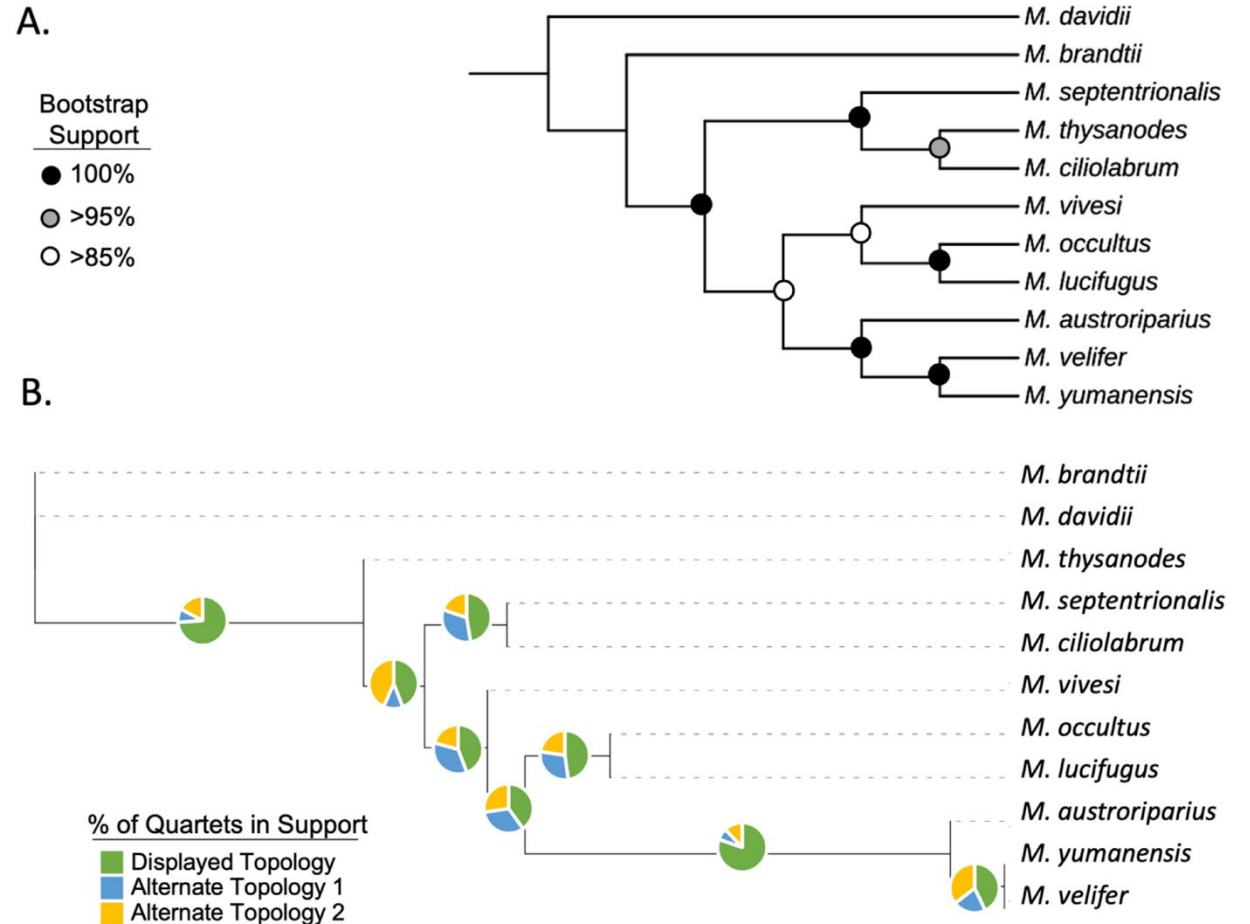
Myotis phylogenomics

- ▣ Massive recent adaptive radiation
- ▣ Phylogeny is unclear



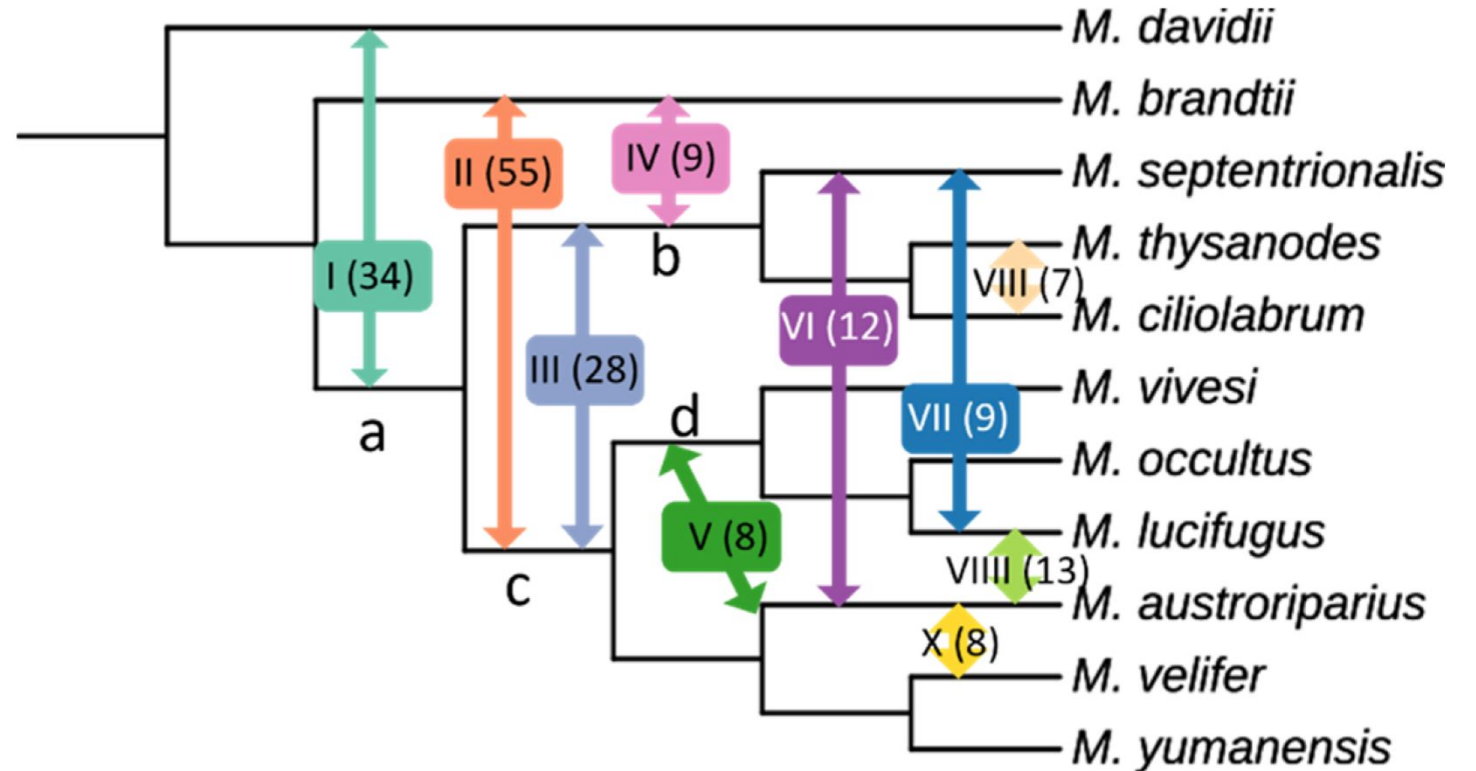
Myotis phylogenomics

- Applying genome-wide markers allowed us to establish a well-supported phylogenetic hypothesis



Myotis phylogenomics

- And identify likely introgression (ancient interbreeding) and/or phylogenetic conflict that help explain the difficulty in understanding the evolution of this group.



Population health and structure - *Myotis septentrionalis* RAD-Seq

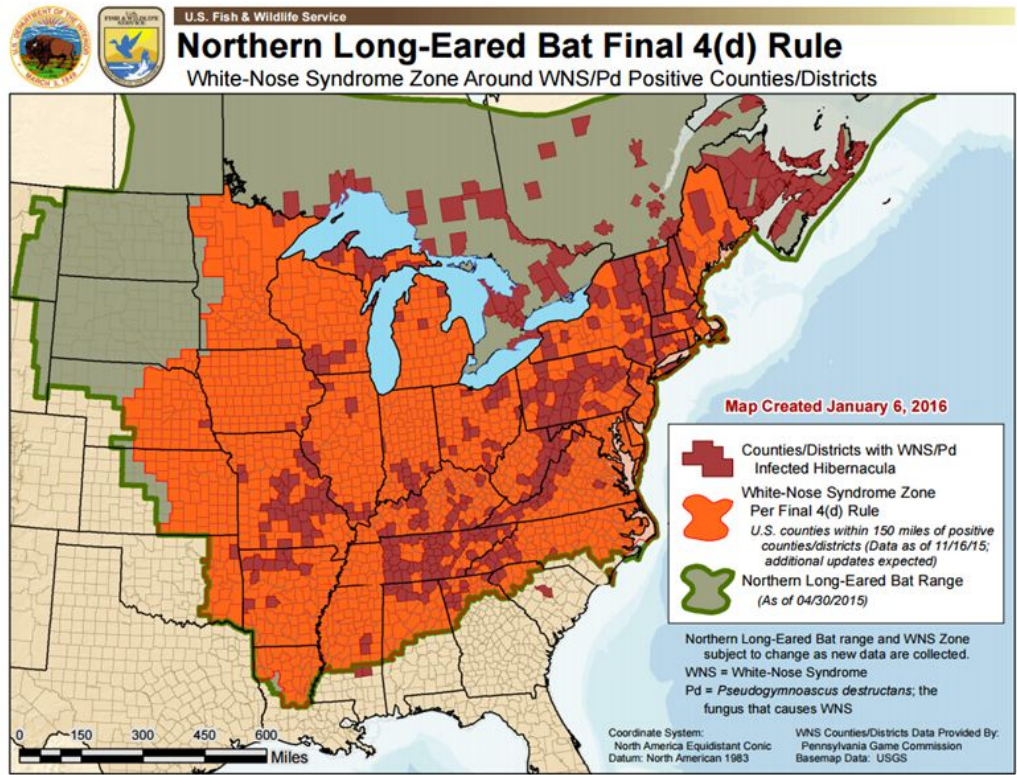
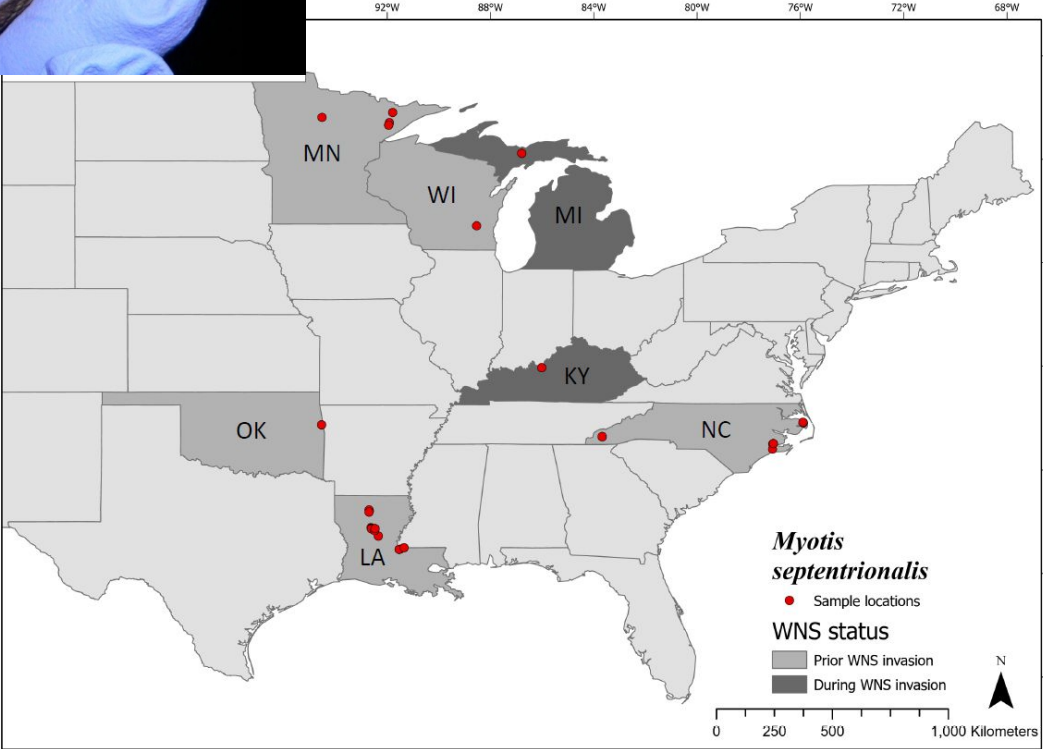
Disentangling genetic diversity of *Myotis septentrionalis*: population structure, demographic history, and effective population size

Jenna R. Grimshaw¹, Deahn Donner², Roger Perry³, W. Mark Ford⁴, Alex Silvis⁵, Carlos J. Garcia⁶, Richard D. Stevens^{7,8}, and David A. Ray^{9*}

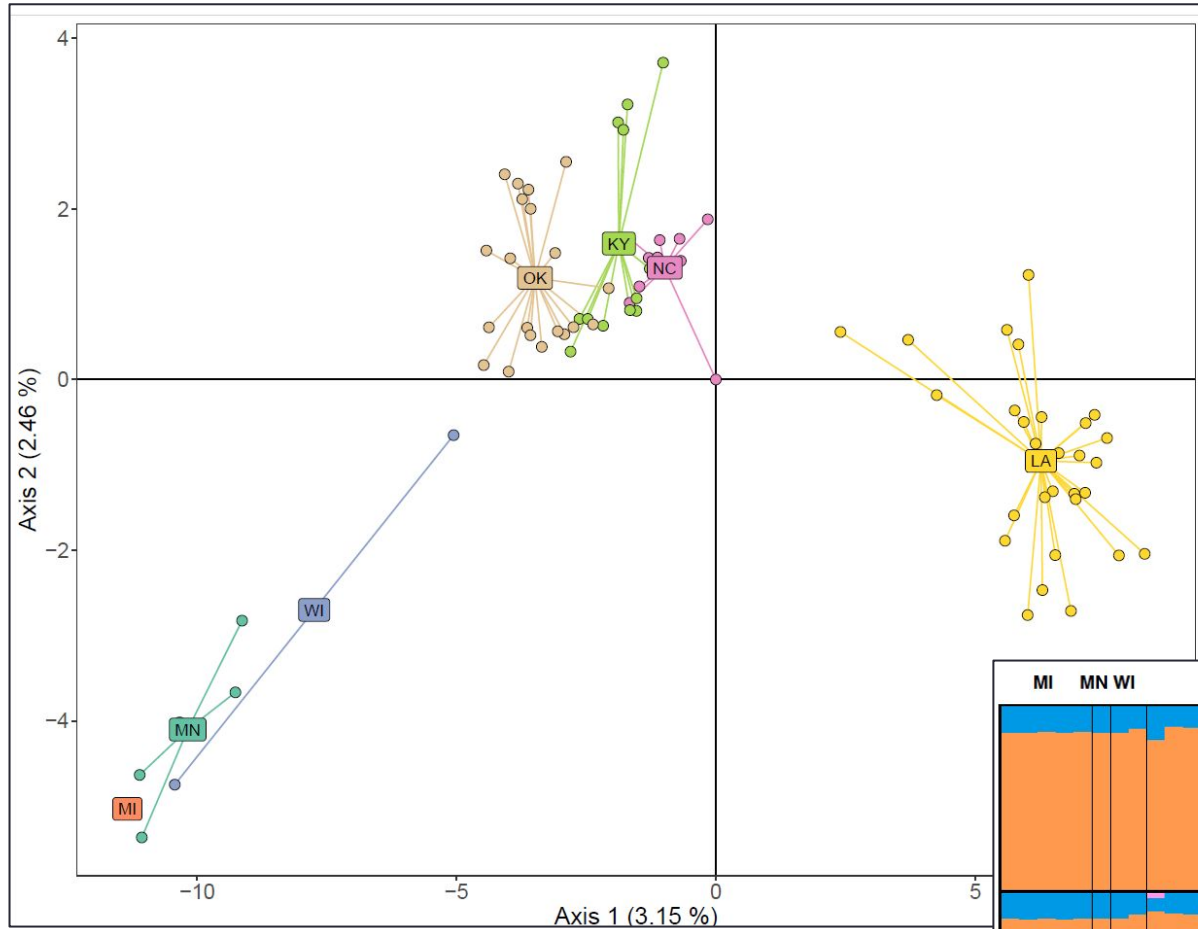


Myotis septentrionalis RAD-Seq

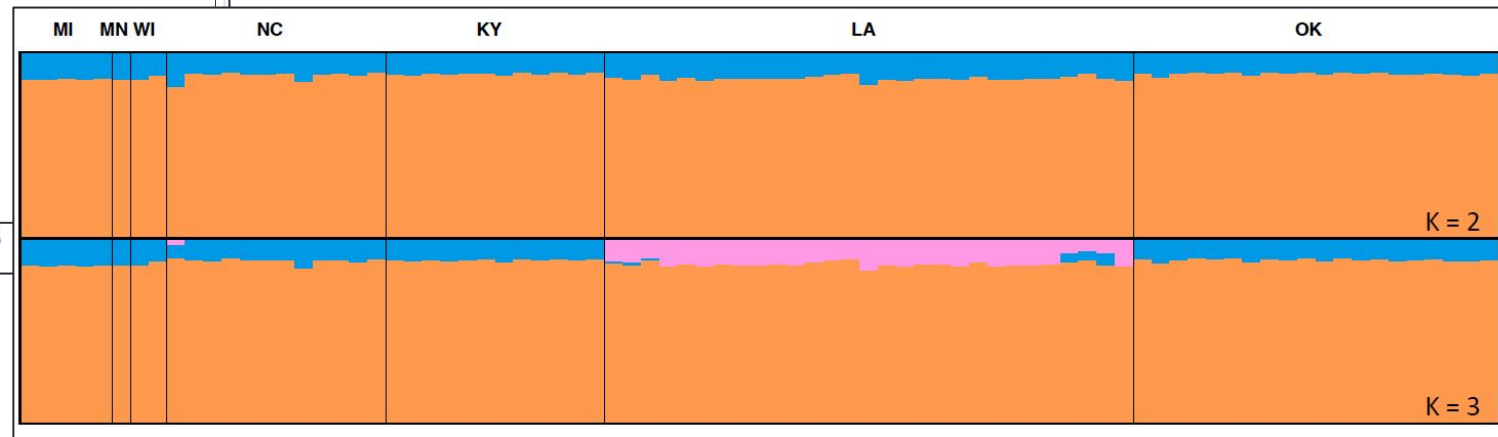
- Widespread but recently hit hard by white-nose syndrome



Myotis septentrionalis RAD-Seq



- Population differentiation is visible, but just barely
- LA is exhibiting signs of genetic uniqueness, likely due to isolation from other populations
- Suggests panmixia in this species
- Data serve as a baseline to investigate impacts of WNS in the future



The potential is vast

Comparative transcriptomics reveals the evidence of genetic adaptations in the macrotis group (Chiroptera: Rhinolophidae)

Insights into the formation and diversification of a novel chiropteran wing membrane from embryonic development

Comparative transcriptome analysis reveals molecular adaptations underlying distinct immunity and inverted resting posture in bats

Single-cell transcriptome analysis of the *in vivo* response to viral infection in the cave nectar bat *Eonycteris spelaea*

The pale spear-nosed bat: A neuromolecular and transgenic model for vocal learning

LANDSCAPE LEVEL PATTERNS OF GENETIC DIVERSITY OF BAT SPECIES OF GREATEST CONSERVATION NEED IN LOUISIANA

Gene losses in the common vampire bat illuminate molecular adaptations to blood feeding

The evolution of antimicrobial peptides in Chiroptera

Recent expansion and adaptive evolution of the carcinoembryonic antigen family in bats of the Yangochiroptera subgroup

Large-scale genome sampling reveals unique immunity and metabolic adaptations in bats

Comparative genomic analyses illuminate the distinct evolution of megabats within Chiroptera

Genome-wide ultraconserved elements resolve phylogenetic relationships and biogeographic history among Neotropical leaf-nosed bats in the genus *Anoura* (Phyllostomidae)

Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats

Inferring genetic structure when there is little: population genetics versus genomics of the threatened bat *Miniopterus schreibersii* across Europe

Take home messages

- Genomes are not the answer to every research question but can be powerful tools to help answer many questions
- Genomes assemblies are difficult to obtain but it's getting easier (and cheaper)
- Genomics is bioinformatically intensive
- High-quality DNA is required for high-quality assemblies
- Applications of genomics to biological problems are varied and informative

Questions?

- david.a.ray@ttu.edu
- <https://www.davidraylab.com/>

